12-2016

Examination and utilization of rare features in text classification of injury narratives

Hsin-Ying Huang
Purdue University

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Hsin-Ying Huang

Entitled

Examination and Utilization of Rare Features in Text Classification of Injury Narratives

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Mark R. Lehto
Chair

Yuehwern Yih

Robert W. Proctor

Victor Raskin

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Mark R. Lehto

Approved by: Abhijit Deshmukh 12/8/2016

Head of the Departmental Graduate Program Date

EXAMINATION AND UTILIZATION OF RARE FEATURES IN TEXT CLASSIFICATION OF INJURY NARRATIVES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Hsin-Ying Huang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2016

Purdue University

West Lafayette, Indiana

DEDICATED TO

My parents, Yi-I and Hsiu-Mei, who raised me into the person I am today
My brother, Wei, who took care of my parents during my absence
My husband, Frank, who believed in "happy wife, happy life"
My in-laws, Yuchang, Tina, and Ray, who warmly welcomed me into their family

Thank you for being a constant source of insight, inspiration, motivation, love, support, and happiness along my life journey.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Mark Lehto, who provided guidance and support with true enthusiasm throughout my doctoral studies at Purdue University. I would like to extend thanks to my advisory committee: Dr. Yuehwern Yih for providing constructive criticism and suggestions, Dr. Robert Proctor for reviewing and improving my dissertation write-up, and Dr. Victor Raskin for supporting my work with the insight from an applied linguistic perspective.

My heartfelt gratitude goes to Dr. Ruey-Yun Horng from National Chiao Tung University in Taiwan for invaluable mentorship and encouragement over the past decade.

Finally, I would also like to thank my colleagues and friends for making life away from home enjoyable. Special thanks are extended to Kerina and Yu-Ling who treat me like family and make me feel at home.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Huang, Hsin-Ying Ph.D., Purdue University, December 2016. Examination and Utilization of Rare Features in Text Classification of Injury Narratives. Major Professor: Mark R. Lehto.

Thanks to the advances in computing and information technology, analyzing injury surveillance data with statistical machine learning methods has grown in popularity, complexity, and quality over recent years. During that same time, researchers have recognized the limitations of statistical text analysis with limited training data. In response to the two primary challenges for statistical text analysis, dimensionality reduction and sparse data, many studies have focused on improving machine learning algorithms. Less research has been done, though, to examine and improve statistical machine learning methods in text classification from a linguistic perspective.

This study addresses this research gap by examining the importance of extreme-frequency words in classifying injury narratives. The results indicate that adhering to the common practice of removing frequently-occurring prepositions from the text significantly decreased the classification performance for certain categories. Removing low-frequency words significantly improved the classification performance for Multinomial Naive Bayes (MNB), helped alleviate the problem of overfitting small categories for Logistical Regression (LR), but did not have any significant effect for Support Vector Machine (SVM).

As a way to utilize low-frequency words, classic word normalization or grouping methods such as stemming and lemmatization are often used in the text preprocessing stage. Despite their popularity, these classic grouping methods are not without limitations. The proposed "Type M+S Word Grouping Method" groups

rare and unseen words morphologically and semantically automatically using unlabeled data. Several experiments were conducted for evaluating the grouping effect for three classifiers (MNB, SVM, LR) in three train-test scenarios (1:9, 1:1, 9:1) on injury surveillance data with a half-million narratives classified into 30 external cause categories. The experimental results show that the proposed method optionally paired with three add-on methods (two-word sequence tagging, reviewed tagging, Naive Bayes-weighted classifier) resulted in better classification performance as compared to stemming and lemmatization. The overall classification performance for small categories with limited training data was improved for MNB (5.5%), SVM (4%), and LR (11.2%) to an extent comparable to increasing the size of the labeled training set by a factor of 3.6 for MNB, 2.3 for SVM, and 5.2 for LR. Some improvement was also observed for medium-sized categories (1.7%) while performance on large categories remained nearly unchanged (0.1%). The overall results advance the conclusion that the proposed method of decision support is a promising approach for incorporating expert knowledge that improves machine learning for classifying injury narratives with reduced manual effort. The results also suggest that simply increasing the size of a training dataset would not result in the level of performance that the proposed method can achieve because of the inherent limitations of linear classifiers to acquire fundamental concepts and classification rules from the narrative that human experts know by definitions of injuries.

1. INTRODUCTION

As valuable information for injury surveillance and prevention, accident description provides richness and depth to the understanding of injury causality (McKenzie, Scott, Campbell, & McClure, 2010). Categorizing or coding injury narratives is part of the analytic process, which is often labor-intensive and error-prone. Due to the advance in information technology and machine learning in the past decade, a growing number of studies have been conducted on utilizing machine learning approaches to classify textual injury data (Lehto, Marucci-Wellman, & Corns, 2009; Lehto & Sorock, 1996; Marucci-Wellman, Lehto, & Corns, 2011, 2015; Nanda, Grattan, Chu, Davis, & Lehto, 2016; Wellman, Lehto, Sorock, & Smith, 2004).

Various empirical evidence has demonstrated the effectiveness of statistical models in analyzing textual data and reducing manual effort in conventional text analysis. While being amazed by the fact that how well simple machine learning and statistical models work, researchers have gradually realized that these automated statistical models can only get to a certain level of accuracy (roughly 70% according to a working paper of Nanda and Lehto where they tested several classifiers on an injury dataset). One of the inherent obstacles lies in the high-dimensionality of textual data, where documents are typically represented as a collection of vectors with the so-called "bag-of-word" or Vector Space Model (VSM) (Salton, 1971). A common form of VSM is the term-by-document matrix, with elements recording the frequency of each word in each document. Unique words in a corpus (vocabulary) often comprise the feature space, as a set of features to represent documents. Consequently, a feature space with a size of tens of thousands words is common in practice, leading to the two primary challenges in statistical text analysis: dimensionality reduction and data sparsity.

Given this high dimensionality of text, there is often a need for dimensionality reduction before applying statistical analysis. While various feature selection methods are available for this purpose, most practitioners choose the so-called Document Frequency Thresholding (DFT) for its simplicity and effectiveness in rapidly reducing feature space. By Zipf (1949)'s law, "given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table" (Chennuru, Chen, Zhu, & Zhang, 2012). The first-ranked (most-frequent) word is expected to have an occurrence frequency that is approximately double of the second-ranked word's frequency and n times of the n-ranked word's frequency. High-frequency words are often stopwords (i.e., function words without semantic meanings but essential grammatically), such as articles and prepositions. Zipf also claimed that the proportion of vocabulary made up by words with a certain frequency f should equal the inverse of $f(f+1)$. In such sense, only half of the vocabulary is retained by just removing the words that only occur once and 20% of the vocabulary is left by removing the words that occur in less than five documents. Zipf's Law justifies the effectiveness of DFT in dimensionality reduction since high- and low-frequency words contribute significantly to the size of vocabulary and total word occurrences in a corpus. Luhn (1958) further advocated the proposition of removing extreme-frequency words by proposing the bell-shaped model that suggests extreme-frequency words have the least discriminatory power. Although DFT has become common practice for text preprocessing and turned out to work adequately well, the question of whether to keep or remove extreme-frequency words in statistical text analysis is still debatable as some empirical evidence has shown that keeping rare words or stopwords can improve the performance of statistical text analysis (Al-Tahrawi, 2013; Mccallum & Nigam, 1998; Pak & Paroubek, 2010; Price & Thelwall, 2005; Riloff, 1995; Saif, Fernandez, He, & Alani, 2014; Schnhofen & Benczr, 2006). Due to the lack of research, this study aims to fill the research gap by providing empirical evidence

on the role of extreme-frequency features in text classification of injury surveillance data.

In addition to a need of dimensionality reduction, the high dimensionality of text also makes statistical models prone to the sparse data problem. Most statistical models have difficulty addressing data rarity especially now in the era of big data. The "curse of dimensionality" is the phrase coined by Bellman (1957) in his mathematical optimization study, which has been widely used in many other domains to describe "various problems that, when analyzing data in high-dimensional spaces, do not occur in low-dimensional settings" (Dirkmaat, 2013). An example of the sparse data problem in an absolute sense is when certain events that do not happen to occur in a training set are mistakenly assigned a probability of zero. Oftentimes, these unseen events compose a great portion of a test set, which negatively influences classification performance. For instance, Essen and Steinbiss (1992) reported that in two text corpora with a 75/25 split, the fraction of word bigrams that occurred in the test set but did not occur in the training set is 12% for the million-word corpus and 50% for the other one-hundred-thousand-word corpus. The problem of unseen events becomes much more serious when annotated training data are limited compared to the availability of unannotated test data, which is often the reality in practice. The other example of the sparse data problem but in a relative sense is that even when certain events occur in the training set, due to their extremely low frequencies, any methods that require statistical significance may still fail to identify these events. Considering a rule association mining problem where two items that rarely occur often occur together when either one is present, the association between them may not be found because random co-occurrences are likely to swamp the meaningful associations between rare items (Liu, Hsu, & Ma, 1999). The sparse data problem has further made statistical models ineffective at identifying rare but meaningful patterns, such as classifying small categories in imbalanced data. The inherent limitation of statistical models lies in their requirement of sufficient training instances to learn

meaningful patterns and explore their relationships. This limitation is unlikely to be fully addressed as the problems of class imbalance and limited availability of annotated data abound in real practice. Thus, this study also plans to examine the potential influence of limited annotated samples, in terms of imbalanced size between categories and between training and test datasets, on the classification performance of various experimental settings.

While the majority of research in relevant fields of text mining has focused on improving algorithms, little research has been done to examine the effectiveness of statistical models in classifying textual data from a linguistics perspective. Feature selection is fundamentally important as it defines the feature space for statistical models to explore and identify patterns. Without a properly-defined feature space, the statistical model is less likely to achieve satisfactory performance. The literature has acknowledged the importance of feature selection in text classification especially for imbalanced dataset (Forman, 2003; Longadge, Dongre, & Malik, 2013). Imagine a machine learning classifier that utilizes a proper feature selection method to capture all discriminative features (i.e., words or concepts that are indicative of certain categories for prediction) and learns from these better-represented training instances. The resulting classification performance is expected to outperform those that do not capture an appropriate feature space in the first place.

A primary reason for the sparsity of the Vector Space Model (VSM) is that words have many synonyms and morphological variants even though they carry similar concepts or share the same hypernym. In addition to removing extreme-frequency words, another common preprocessing task is to apply a word normalization or grouping method, such as stemming or lemmatization. These classic grouping methods group words with similar spelling by removing the ending of words and merging them with their root or base form. The underlying assumption is that words with the same root or base form carry similar or related semantic meaning. Despite their popularity, stemming and lemmatization have limitations in

addressing misspellings, domain-specific words, multi-word sequences, and synonyms with dissimilar spelling. Thus, the other objective of this study is to propose the so-called Type M+S Word Grouping Method that aims to address some of the limitations of stemming and lemmatization by grouping rare and unseen (words only occurred in test set) words by their linguistic features (morphology and semantics). As the literature has acknowledged the potential of statistical semantics in capturing the semantics of human natural language (Baroni, Dinu, & Kruszewski, 2014; Han, Cook, & Baldwin, 2012; Pecina, 2005), this study plans to explore the feasibility of using statistical semantics in identifying same-hypernym words for discriminative concepts. The existing statistical methods are utilized in a mix-and-match fashion to minimize human effort and automate the grouping process while allowing for optional incorporation of human review. By creating a smaller but denser VSM with a set of more statistically robust, discriminatory features, the proposed Grouping Methods (optionally paired with add-on methods: reviewed tagging, word-sequence tagging, and applying Naive-Bayes weighted classifiers) are expected to improve standard classification without grouping and even to perform more effectively in the case of limited annotated samples (i.e., in small categories in an imbalanced dataset or in a small training dataset relative to the test dataset size).

1.1 Objectives

The overall goal of this study is to improve the limitations of existing statistical machine learning approaches to address data rarity and imbalance (limited, imbalanced labeled samples between classes or training and test datasets) from a linguistic perspective.

The first objective of this study is to examine the role of extreme-frequency words in text classification of injury narratives. The extreme-frequency features

are considered important and should not be removed arbitrarily if their absence can negatively impact classification performance.

The other objective of this study is to propose the Word Grouping Method of utilizing both rare and unseen words for better classification performance by grouping them by linguistic features, morphologically and semantically. This study reviews and examines the feasibility of using statistical semantics in identifying same-hypernym words for grouping, along with several statistical methods for nonconventional purposes: using the coefficient matrix of a linear classifier to identify the predictive category for words and using feature selection methods to identify discriminatory features for grouping. The proposed method aims to outperform the classic word grouping methods of stemming and lemmatization by addressing some of their limitations and even to perform more effectively when annotated samples are limited (i.e., small categories or training dataset).

1.2 Organization

The remaining chapters are organized as follows:

Chapter 2: **Literature Review** reviews relevant research on two main areas close to the research interest of this study: Vector Space Model (VSM) and challenges for statistical text analysis. First I start with the introduction of VSMs and its applications, including how to statistically measure word similarity and construct VSMs, followed by the overview of two types of word similarity and their measures, distributional semantic models and Google's Word2Vec model, and three classic classifiers used in this study. Finally, two primary challenges of high-dimensionality of text for statistical text analysis are discussed, including feature selection and data sparsity.

Chapter 3: **Research Objectives and Method** first describes the research objectives and hypotheses of this study. This chapter then introduces the dataset used in this study, along with the list and frequency distribution of external cause

categories for prediction, followed by the method and performance evaluation for classification tasks.

Chapter 4: **Role of High- and Low-Frequency Words in Text Classification** is dedicated to accomplishing the first objective of this study by conducting the experiments to explore the importance of frequently-occurring stopwords and low-frequency words in text classification of injury narratives, specifically the impact on the classification performance due to their removal.

Chapter 5: **Introduction of Type M+S Grouping Method** first discusses the background and limitations of stemming and lemmatization. The chapter then introduces the so-called Type M+S Grouping Method that aims to address some of the highlighted limitations and provide an overview of the Type-M Morphological Mapping and Type-S Semantic Grouping methods in the following chapters.

Chapter 6: **Type-M Morphological Mapping Method** describes the chapter's namesake method in more detail, followed by the experimental design. The effect of Type-M Mapping, along with the feasibility of using the coefficient matrix of three classic classifiers in identifying predictive categories for words, are then examined, discussed, and summarized in the end.

Chapter 7: **Type-S Semantic Grouping Method** continues with a more detailed introduction to the proposed Semantic grouping. First, an overview of the two main branches of statistical semantics (i.e., correlational and distributional semantics) is given, followed by a feasibility study to compare these two types of statistical semantics in identifying same-hypernym (similar-concept) words. Next, I further explore the superior statistical semantic measure, Word2Vec, identified in the experiment by starting with an exploratory study to gain insight into the mechanism of Word2Vec in quantifying word similarity and ranking similar words. Given the better understanding of how Word2Vec works, the Type-S Semantic Grouping Method that utilizes Word2Vec is introduced to group same-hypernym words with dissimilar spellings. Experiments are then conducted to investigate the effectiveness of Semantic Grouping in improving classification performance,

along with two major factors that can influence the performance: grouping strategy (tagging or mapping) and manual review effort (combined factor of threshold and review level). Lastly, in order to improve the scalability of the Type-S Semantic Tagging Method for the data with a large category set, I also explore the feasibility of pairing Semantic Tagging with classic feature selection methods to prioritize discriminatory words and automatically identify seedwords for semantic grouping to improve the classification performance without any human effort.

Chapter 8: **Study Summary and Final Evaluation** first summarizes the experimental results that have been presented so far, evaluates the proposed Type M+S Grouping Method and three promising add-on methods for further improvement, benchmarks the proposed methods with the classic word grouping methods of stemming and lemmatization, highlights the most significant findings of this research, and concludes by discussing some of the most promising future research that can follow this study.

2. LITERATURE REVIEW

2.1 Vector Space Model (VSM) of Semantics

2.1.1 Introduction to VSM

One of the largest impediments to human computer interaction is that computers do not understand human language. This, in turn, greatly limits computers to process and analyze unstructured data of natural language. The Vector Space Model (VSM) developed by Salton and his colleagues in the 1970s has made a pioneering attempt to address these limits and shown promising results in the relevant fields of natural language processing and text mining (Salton, Wong, & Yang, 1975; Salton, 1971). The basic idea of the VSM is to represent a corpus of documents as a collection of word vectors. The elements in the vector are derived from event frequencies, for example: how often a given word is presented in a given context such as a document, a sentence, or a window of words. Thus, the VSM is also known as the "bag-of-words" method as it treats text like "a bag of words".

The emphasis on event frequencies explicitly connects VSM to Harris' (1954) distributional hypothesis in linguistics, derived from the semantic theory of language usage. VSMs operate under the distributional hypothesis, assuming that "words that occur in similar contexts (i.e., distributionally similar words) tend to have similar meanings" (Harris, 1954). Around the same time, Weaver (1955) suggested to disambiguate word senses by looking at the words around an ambiguous words in a window of N content words, where N depends on the statistical semantic character of language. Later, Firth (1957) popularized this idea with his notion of "you shall know a word by the company it keeps!" and now the distribution hypothesis has been extended to several hypotheses for different application

purposes (see section 2.7 Hypotheses in Turney & Pantel (2010) for more details). By providing an efficient and quite reliable way of extracting knowledge automatically from a given corpus and measuring and quantifying the similarity of meaning between words or documents, VSMs have been used in most search engines for identifying documents associated with a user-specified query (Manning, Raghavan, & Schtze, 2008) and also in many leading algorithms for measuring semantic relatedness (Pantel & Lin, 2002; Rapp, 2003; Turney, Littman, Bigham, & Shnayder, 2003). Thus, the invention of VSM has made the statistical semantics possible. Although Weaver (1955) was the one who first used the term, Delavenay (1960) formally defined statistical semantics as "statistical study of meanings of words and their frequency and order of recurrence."

The most popular representation of VSMs is the "term-document matrix." Given a corpus containing "a total of d documents described by t terms, the term-document matrix A is a $t \times d$ matrix, where the columns of A are document vectors and the rows of A are term vectors. The matrix element a_{ij} is the weighted frequency at which term t_i occurs in document d_j" (Berry, Drmac, & Jessup, 1999). Several term weighting methods have been proposed such as TF-iDF (Term Frequency-Inverse Document Frequency) (Sparck Jones, 1972). Term-document matrices were first developed for document retrieval by calculating the semantic similarity between documents and queries and retrieving the top-ranked relevant documents. The relevance between documents and queries (which can also be treated as documents) is given by the similarity of their vectors using a similarity measure, such as cosine similarity. In general, the cosine measures the orientation of two vectors: two vectors with the same and opposite orientation have a cosine similarity of 1 and -1, while two vectors at 90 degree have a similarity of 0. Cosine similarity has an output bounded in [0, 1] in a positive space with increasing level of similarity. This feature is utilized in the fields of information retrieval and text mining, where each word or document is characterized by a vector using the VSM model. Featuring the ability of measuring similarity between documents, term-

document matrices have been extensively used in text clustering or classification, question answering, essay grading, document segmentation, and call routing (see the section of Applications in Turney & Pantel (2010) for details).

Motivated by the term-document matrix of Salton et al. (1975), Deerwester and his colleagues shifted from measuring the similarity of column vectors (documents) to measuring the similarity of row vectors (words) (Deerwester et al., 1990). The distributional hypothesis serves as the justification for applying the VSM to measuring word similarity, stating that "words that occur in similar contexts tend to have similar meanings" (Firth, 1957; Harris, 1954). The so-called "word-content matrix" is then proposed to represent a word by a context vector, in which the elements are derived from the occurrences of the words in various contexts, such as sentences, paragraphs, documents (same as term-document matrix), windows of words (Lund & Burgess, 1996), or grammatical dependencies (Lin, 1998). Similar to measuring the document frequency, the word similarity can be measured by the cosine of their context vectors. Two words are said to be similar if their context vectors are distributionally similar, which happens when the cosine similarity is close to 1. Two words are said to be independent and dissimilar if the value is close to zero. In addition to word similarity, other popular application areas include word clustering and classification, automatic thesaurus generation, word sense disambiguation, context-sensitive spelling correction, semantic role labeling, query expansion, textual advertising, and information extraction (see the section of Applications in Turney & Pantel (2010) for details).

2.1.2 Construction of VSM

The construction of VSMs involves a series of linguistic and mathematical processing tasks. Turney & Pantel (2010) presented a comprehensive review on VSMs and the processing tasks involved in the construction. The key steps are highlighted in the following:

1. Linguistic processing

   (a) Tokenization: to tokenize or segment raw text by determining what constitutes a term (single-word or multi-gram terms) and how to extract terms from raw text using proper delimiters.

   (b) Stop-word removal: stop-words are the words considered as non-descriptive within bag-of-words approach, typically prepositions, articles, etc.

   (c) Normalization: since different strings may have the same meaning (synonymy), one can normalize superficial variations by merging them into their base form. Common normalization methods include case folding, stemming, lemmatization, or self-defined controlled vocabulary.

   (d) Annotation: the inverse of normalization. Since a string may have different meanings depending on the context (polysemy), one can identify and disambiguate them with proper marking. Common forms of annotation include word sense tagging, part-of-speech tagging, and parsing (syntactic analysis that analyzes the grammatical structure of sentences to identify their grammatical roles).

2. Mathematical processing

   (a) Building the frequency matrix

   "An element in a frequency matrix corresponds to an event: a certain item (term, word, word pair) occurred in a certain situation (document, context, pattern) a certain number of times (frequency)" (Turney & Pantel, 2010).

   (b) Weighting the elements

   The elements in the VSM represent how important or discriminant a term is in a set of documents. In general, a weighting scheme is designed to weigh less on common events but weigh more on unexpected events.

The most popular weighting scheme is the Term Frequency — inverse Document Frequency (TD-iDF) family (Salton & Buckley, 1988).

(c) Smoothing the matrix for dimensionality reduction

One well-known drawback of the VSM is its high dimensionality, and smoothing the matrix by reducing the dimensions may be the most common solution. "Latent semantic indexing" (LSI) proposed by Deerwester et al. (1990) was a pioneering success to improve similarity measurements between documents. The LSI utilizes singular value decomposition (SVD) to represents both terms and documents as vectors in a space of self-selectable (usually lower than the original) dimensionality. In this reduced model (truncated SVD or thin SVD), the similarity measurement using the dot product or cosine between documents or terms is expected to be improved as the similarity is now approximated by values on a smaller number of dimensions, which are considered to be a set of uncorrelated indexing variables or factors (latent meaning). In addition to the LSI for improving document similarity, "Latent Semantic Analysis"(LSA) is for improving word similarity using truncated SVD. Landauer & Dutnais (1997) applied truncated SVD to word similarity and achieved human-level scores on multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL).

As the name suggests, "Latent Semantic Indexing or Analysis" is a method for discovering latent meanings (Deerwester et al., 1990; Landauer & Dutnais, 1997). The low-dimensional mapping is expected to capture the latent meaning in the words and the contexts. "Limiting the number of latent dimensions forces a greater correspondence between words and contexts," which is believed to improve the similarity measurement (Turney & Pantel, 2010).

2.1.3 Word Similarity

According to Saussure (1916), word similarity can be classified as either syntagmatic similarity or paradigmatic similarity. The Syntagmatic relation concerns

position, which is also referred to as first-order (1-st) relation, attributional similarity, semantic relatedness, or correlational similarity (Han et. al., 2013). Two words are syntagmatically related if they co-occur in a given text more frequently than chance. The reason is that the syntagmatic associates often involve a certain type of semantic relation to the degree that they share attributes, and thus they are more likely to be mentioned together in a specific context (Turney, 2006). Thus, syntagmatic associates tend to be the neighbor of each other and they often have different part-of-speech tags. Some modified examples from Turney & Pantel (2010) for different semantic relations are synonyms ("bank" and "trust company"), meronyms ("car" and "wheel"), antonyms ("hot" and "cold"), and words that are functionally related ("knife" and "cut") or frequently associated ("pencil" and "paper"). The Paradigmatic relation, on the other hand, concerns substitution. They are also called second-order (2-nd) relation, semantic similarity or taxonomical similarity (Turney & Pantel, 2010). Paradigmatic associates tend to have similar neighbors and they are often substitutable for one another in a specific context (e.g. "doctor" and "nurse", "apple" and "orange"). Since the two paradigmatic-related words are a good substitute for each other, they are likely to either be synonyms / antonyms or share a hypernym. For instance, with a common hypernym, "doctor" and "nurse" are similar as they are medical personnel and "apple" and "orange" are fruits.

Since words are said to have a syntagmatic relation if they co-occur more frequently than chance, the measure of syntagmatic relation can be derived from their co-occurrence information (Sahlgren, 2006). Some of the well-known association measures for statistical dependency include the pointwise mutual information (PMI), t-test, Pearson's Chi-square test, and log-likelihood ratio. Pecina (2005) compared 84 association measures for collocation extraction and found the PMI to be superior to other measures.

On the other hand, the 2-nd order relation, also known as paradigmatic association or distributional relation, relates words that share similar neighbors. Thus,

the paradigmatic similarity of two words can be determined by the agreement of their lexical neighborhoods (Ruge, 1992), which is often measured by the cosine of the two context vectors. Rapp (2002) termed the method that relies on the word co-occurrence information as the 1-st order approach and the one that utilizes context vectors as the 2-nd order approach. He compared these two approaches and found that 2-nd order approach produces paradigmatic associates (e.g. blue: *red*, *green*, *grey*, *yellow*, *white*) and the 1-st order approach produces mixed results (e.g. blue: *red*, *eyes*, *sky*, *white*, *green*). As a result, Rapp (2002) suggested to combine both approaches to improve the results of finding semantically associated word-pairs. Table 2.1 below summarizes the comparison between the first-order correlaitonal relation and second-order distributional relation.

Table 2.1: Comparison between Syntagmatic and Paradigmatic Relation

	Correlational Semantics	**Distributional Semantics**
Other names	First-order relation; Syntagmatic relation; semantic relatedness; attributional similarity	Second-order relation; Paradigmatic relation; semantic similarity; taxonomical similarity
Definitions	Two words co-occur more frequently than by chance and they tend to be the neighbor of each other	Two words tend to have similar neighbors and they are substitutable in a specific context
Tendency of Semantic	Relatively loose semantic relations. Any sematic relations, including synonyms, meronyms, antonyms, or functionally related or frequently associated.	Relatively tight semantic relation. Often synonyms / antonyms or share a hypernym.
Examples	Bank-trust company; car-wheel; hot-cold; knife-cut; pencil-paper	doctor-nurse; apple-orange
Measurement basis	Occurrence and co-occurrence frequencies	Context vectors
Measurement methods	first order approach that relies on co-occurrence information, including: pointwise mutual information (PMI), t-test, Pearson's Chi-square test, and log-likelihood ratio	Second order approach that relies on context vectors, including: Cosine, city-block distance
Extracted examples (Rapp, 2002)	blue: red*, eyes, sky, white*, green* (*including the paradigmatic associates)	blue: red*, green*, grey, yellow, white*

2.1.4 Distributional Semantic Model and Google's Word2Vec Model

In computational linguistics, it is widely believed that words that occur in the same context tend to share similar meaning. This idea was first proposed by Harris

in 1954, later popularized by Firth (1957)'s saying: "a word is characterized by the company it keeps." This concept is the so-called "distributional hypothesis," where words can be quantified based on their distributional properties in large samples of text data and words tend to share similar meaning if they have similar contextual distributions (i.e., word vectors or representations in the vector space model).

In the research area of distributional semantics, the distributional semantic models (DSM), also known as "word space" model, are used for quantifying the semantic similarity between language units under the distributional hypothesis. Such quantification method has been widely applied to a variety of linguistic tasks such as solving the TOEFL synonym test (Landauer & Dumais, 1997; Rapp, 2004), identification of translation equivalents (Rapp, 1999), word sense induction and discrimination (Schutze, 1998), POS induction (Schutze, 1995), identification of analogical relations (Turney, 2006), semantic classification (Versley, 2008), and so on.

DSMs can be classified into two groups depending on the nature of learning: Count-based models (unsupervised) and Predictive-based models (supervised) (Baroni et al., 2014). Count-based models are traditional DSMs, which focus on the reweighting or transformation of the original matrix with elements of word co-occurrence counts. This process is generally unsupervised and often involves matrix algebra for normalization or dimensionality reduction using singular value decomposition. Examples of classic count-based models include Pointwise Mutual Information and Latent semantic analysis. Predictive-based models are more recent development of DSMs, which are also called "neural probabilistic language models" because they are trained by neural network for supervised learning. The text data is framed as a supervised task without involving any manual annotation. Specifically, the model is trained to predict the contexts given a target word, or vice versa (see Skip-gram model in later section). The weights in the word vectors can then be optimized using techniques such as stochastic gradient descent and back-

propagation. According to the distributional hypothesis, a properly trained model should learn to assign similar context vectors to words with similar meaning.

Both Count-based and Predictive-based models can be considered as word embedding techniques if original word vectors can be represented in a lower dimensionality through certain mapping structures. Count-based DSMs rely on the statistics based on the word co-occurrence matrix and then try to normalize and map these count-statistics down to a set of small, dense word vectors. Predictive-based DSMs, on the other hand, aim to predict a word from its context (or vice versa) with learned small and dense word vectors. In contrast to the more heuristic nature of count-based DSMs, predict-based DSMs seems to be more grounded with a well-defined supervised learning step. More detailed comparison between these two models is elaborated in Baroni et al. (2014)'s paper, where the authors systematically compared and evaluated both models and their finding concluded that predict-based DSMs are indeed superior to count-based DSMs on syntactic and semantic tasks.

As a type of predictive-based or neural probabilistic language model, the Word2Vec model has attracted a great amount of attention and research interests recently. The model, developed by Tomas Mikolov and colleagues in 2013, seems to be capable of capturing the linguistic features of human languages, both syntactically and semantically. A properly trained Word2Vec model can figure out the semantic relations such as capital city, currency, and gender, or syntactic relations such as word tense, singular/plural, opposite, and comparative or superlative. Some interesting findings from Mikolov, Yih, and Zweig (2013) are provided below. An example of word pairs with a syntactic "superlative" relationship can be big – biggest and small – smallest. Given the word pair "biggest – big" and the word "small", the model can suggest a word (i.e., "smallest") that is similar to "small" in the same sense as "biggest" to "big". Another analogical example of semantic relationship "city and the country it belongs to" can be "Paris" is to "France" as "Rome" to "Italy". Answers to such analogical questions can be found by per-

forming simple algebraic operations with the relevant word vectors learned by the model. Using the syntactic example above, the answer is the word whose word vector is the closest to the resulting vector X = vector("biggest") – vector("big") + vector("small").

Although being marketed as the deep learning technique on the website of Python's Gensim library, Word2Vec should be considered as a word embedding method because its trained neural net is quite shallow. Word2Vec can be viewed as a simpler, but more computational efficient version of neural network language model (NNLM). Word2Vec utilizes similar training mechanisms as NNLM (i.e., stochastic gradient descent and backpropagation) but its simpler model architectures allow it to train on a great amount of data with much less time and to achieve comparable or slightly better performance on semantic-syntactic analogical relation tasks. The following sections briefly discuss these two model architectures: Continuous bag-of-words (CBOW) and Skip-gram Model. Refer to Rong (2016) for a detailed mathematical explanation for Word2Vec models.

- **Continuous Bag-of-Word Model (CBOW)**

CBOW is used to predict the word in the middle of its context words within a symmetric window based on the sum or mean of these context word vectors.

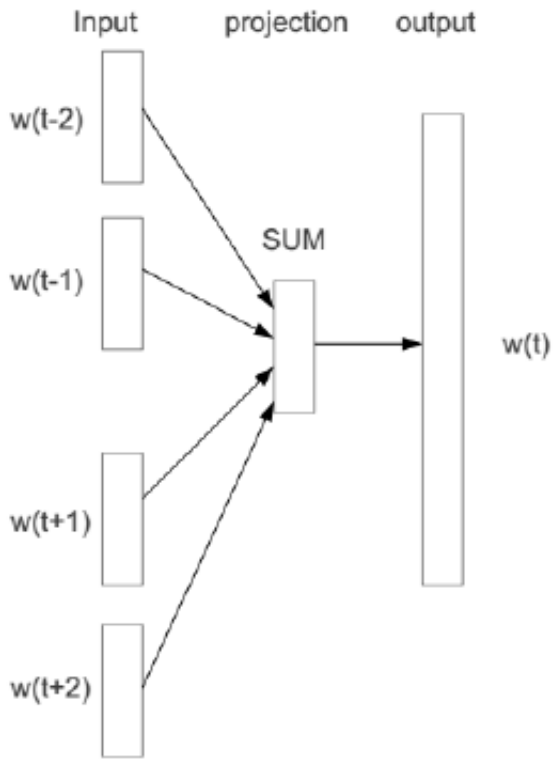

Figure 2.1: Explanation Example of Continuous Bag-of-Word Model

For example, given a text narrative: "PAIN LEFT ARM POST FALL AT SCHOOL YESTERDAY." Given a window of two, (i.e., two words are the maximum distance between the current word and predicted word within a sentence), we have a list of (context, target) word pairs in the following, used as (input, output) pairs for training to predict a target word given context words:
([PAIN, LEFT, POST, FALL], ARM), ([LEFT, ARM, FALL, AT], POST), ([ARM, POST, AT, SCHOOL], FALL),...

The objective function is to maximize the conditional probability of observing the actual target word W_T given the context word w_C using soft-max function:

$$P(w_T|w_c) = \frac{exp(v'_{wT} \cdot v'_{wc})}{\sum_{j=1}^{v} exp(v'_{w'j} \cdot v'_{wc})} \tag{2.1}$$

where v_w and v'_w are two vector representations of the word w . v_w is called input vector, coming from the weighting matrix W from input layer to hidden layer, while v'_w is called output vector, coming from the other weighting matrix W' from hidden layer to output layer. V is the vocabulary size. For a multi-word context setting, v_{wC} can be the sum or mean of the input vectors of context words to optimize vector representations of words. The CBOW model is trained by maximizing the log-likelihood function of $P(w_T \mid w_c)$:

$$max\, P(w_T|w_c) = max\, logP(w_T|w_c) = v'_{wT} \cdot v'_{wc} - log(\sum_{j=1}^{v} exp(v'_{w'j} \cdot v'_{wc}) \qquad (2.2)$$

- **Skip-gram Model**

The Skip-gram model does the inverse of the CBOW model. The target word is now at the input layer and the context words are at the output layer. The Skip-gram model is trained to predict the context given the target word. As its name "skip" tells, the context is not limited to the target word's immediately adjacent words, i.e., context words can be $w(t \pm j), j \geq 1$

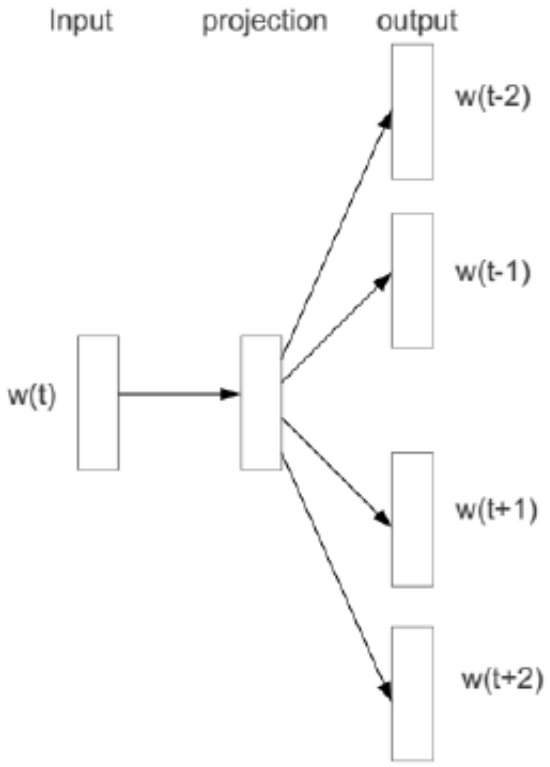

Figure 2.2: Explanation Example of Skip-gram Model

Now the context and target are inversed for (input, output) pairs. The Skip-gram model is trained to predict each context word from its target word. Using the narrative example earlier, the (input, output) pairs for training become: (ARM, PAIN), (ARM, LEFT), (ARM, POST), (ARM, FALL), (POST, LEFT), (POST, ARM), (POST, FALL)...

Similar to the CBOW model, the objective function of the skip-gram model is to maximize a set of C conditional probabilities of actually observing the c-th context word $w_{C,c}$ given the target word w_T using soft-max:

$$P(w_{C,c}|w_T) = \frac{exp(v'_{wc} \cdot v'_{wT})}{\sum_{j=1}^{v} exp(v'_{w'j} \cdot v'_{wT})} \tag{2.3}$$

where $w_{C,c}$ is the c-th context word for $c = 1, 2, \cdots, C$.

The log-likelihood function to be maximized now becomes:

$$\begin{aligned} max\, P(w_{C,1}, w_{C,2}, ..., w_{C,C}|w_T) &= max\, log(\prod_{i}^{C} P(W_{C,i}|W_T)) \\ &= max \sum_{i}^{C} log P(W_{C,i}|W_T) \\ &= \sum_{i}^{C} V'_{w_{C,i}} \cdot V_{wT} - C\, log(\sum_{j=1}^{V} exp(V'_{w',j} \cdot V_{wT})) \end{aligned} \tag{2.4}$$

- **Optimization of Word2Vec and scaling up for large corpus**

With the defined objective functions in Equations 2.2 and 2.4, stochastic gradient descent and backpropagation are then applied for training and optimization. The performance of both models is similar if they are trained for sufficient number of epochs, but the CBOW model is relatively more computational efficient so it is recommended for learning a larger corpus (Mikolov et al., 2013). Recall that, in both models, every word in the vocabulary has two vector representations: the input vector V_w and the output vector V'_w. Learning the output vector is much more computationally expensive than learning the input vector. To update the output vector, for each training instance, we need to calculate $\sum_{j=1}^{V} exp(V'_{w',j} \cdot V_{wc}$ for the CBOW model (or $\sum_{j=1}^{V} exp(V'_{w',j} \cdot V_{wT}$ for skip-gram). More precisely, for each training instance, we need to iterate through every output word (i.e., the entire vocabularies) to compute the dot product of its output vector $V'_{w',j}$ and the input vector of current input word V_wc or V_wT, probability prediction, prediction error, and finally update the output vector using the prediction error. Going through such computations for every word in all training instances is impractical to scale up to a large training corpus or vocabularies. The intuitive solution is to limit the number of output vectors ($V'_{w',j}$ for $j = 1, 2, \cdots, V$) that must be updated for each training instance.

Word2Vec offers two optimization tricks that optimize the repetitive computations for updating output vectors: Hierarchical softmax and Negative Sampling. Mikolov et. al. (2013) and Rong (2016) have detailed explanations, but this study summarizes the basic principles in the following.

Hierarchical softmax, first introduced by Morin and Bengio in 2005, was able to approximate of the full softmax with improved computational efficiency. Hierarchical softmax utilizes a binary tree (learned using the Huffman tree generator) to represent all words in the output layer, where words are leaf units and nodes represent the relative probability of their child nodes. For each leaf unit, there exists a path from the root (in the hidden layer) to the unit (in the output layer). By defining a random walk that assigns the probability to each word from the root in the hidden layer to the leaf unit in the output layer, the number of output vectors to evaluate for each training instance has significantly reduced from V (the entire vocabularies) to about $log_2(V)$. Figure 2.3 shows a unique path for W_2 from the root to the leaf unit, where $n(w,i)$ is the i-th unit on the path from the root to the word w.

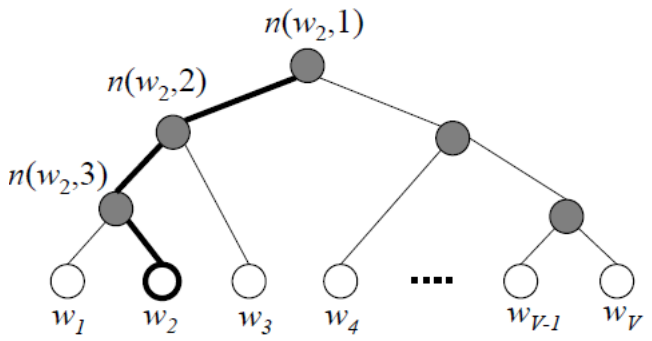

Figure 2.3: Explanation Example of Softmax Model

Without involving mathematical optimization equations, the probability of W_2 being the output word W_0 can be understood as:

$$P(w_2 = w_0) = p(n(w_2,1), left) \cdot p(n(w_2,2), left) \cdot p(n(w_2,3), right) \tag{2.5}$$

An alternative to hierarchical softmax is Negative Sampling, a simplified version of Noise Contrastive Estimation, introduced by Gutmann and Hyvrinen in 2012. The idea of Negative Sampling is rather more straightforward: it estimates the probability of an output word by learning to distinguish it from draws (as negative samples) from a noise distribution. For example, suppose the word "overdose" appears along with the context word "drug", then the vector of "drug" should be more similar to the vector of "overdose" (as measured by their dot product) than the vectors of several other randomly chosen words from a probabilistic distribution, e.g. "struck", "electrocution", "floor". The probabilistic noise distribution can be determined arbitrarily or empirically. Word2Vec uses the unigram distribution raised to the 3/4th power as the noise distribution, as Mikolov et al. (2013) found it significantly outperformed the unigram and the uniform distributions on all their experiments. The Word2Vec toolkit suggests 5 to 20 negative samples drawn from the noise distribution.

The objective function, for each instance, is to maximize the probabilities of seeing that the target word and context word indeed came from the data and that each noise word and context words did not occur together:

$$\begin{aligned} & max\, P(D=1|w_T, w_C) \prod_{i}^{k} P(D=0|\widetilde{(w_i)}, w_c) \\ = & max\, logP(D=1|w_T, w_C) \prod_{i}^{k} P(D=0|\widetilde{(w_i)}, w_c) \\ = & log Q_\theta(D=1|w_T, w_C) + \sum_{i=1}^{k} log Q_\theta(D=0|\widetilde{(w_i)}, w_c) \end{aligned} \tag{2.6}$$

where $Q_\theta(D=1|w_T, w_C)$ is the binary logistic regression probability under the model of seeing that the target word w_T occurs with the context word w_c in the dataset D. k noise words were drawn from the noise distribution as negative examples.

In addition to applying the hierarchical softmax or negative sampling in the training process, the computational efficiency can also be improved through reducing the vocabulary size in the data pre-processing step prior to the training. One method is to arbitrarily set a frequency cut-off that excludes the words whose document frequency is below this threshold (i.e., the parameter "min_count" in Word2Vec function), or to set a maximum number of vocabulary size (i.e., the parameter "max_vocab_size"). Another method implemented in Word2Vec is through subsampling of high-frequency words. Word2Vec subsamples and randomly discards the frequent words whose frequency is higher than a threshold t, with a probability $P_{discard}$ that is proportional to their frequencies.

$$P_{discard}(W_i) = 1 - \sqrt{\frac{t}{f(w_i)}} \propto f(w_i) \tag{2.7}$$

The underlying reason is that high-frequency words are often stopwords (e.g. "the", "in", and "a"), which are considered as less informative than rare words. Also, the vector representations of high-frequency words usually do not change significantly after training on several million examples. This aggressive way of diluting high-frequency words was found to improve not only the training speed but also the quality of resulting learned word vectors of the rare words (Mikolov et al., 2013).

2.1.5 Machine Learning Classifiers

This section provides a brief introduction to three classic linear classifiers used in this study: Naive Bayes, Support Vector Machine, and Logistic Regression.

- **Naive Bayes (NB)**

The Naive Bayes classifier is one of the most commonly used and simplest classifiers (Mccallum & Nigam, 1998). The predicted probability of a category C_i given

a set of n words in a document is defined as:

$$P(C_i|n) = \prod_{j=1}^{n} \frac{P(n_j|C_i)P(C_i)}{P(n_j)} \tag{2.8}$$

where $P(C_i)$ is the prior probability of a category C_i, $P(n_j)$ is the prior probability of word n_j, and $P(n_j|C_i)$ is the probability of the word n_j given the category C_i. $P(C_i)$, $P(n_j)$, and $P(n_j|C_i)$ are estimated based on their frequency in a training set. In practice, the Laplace smoothing or additive smoothing is often applied to$P(n_j|C_i)$ by adding a small constant α to the number of times a particular word occurred in a category. That is:

$$P(n_j|C_i) = \frac{count(n_j|C_i) + \alpha \times count(n_j)}{count(C_i) + \alpha \times N} \tag{2.9}$$

where $count(n_j|C_i)$ is the number of times a word n_j occurs in category C_i, $count(n_j)$ is the number of times a word n_j occurs, $count(C_i)$ is the number of times a category C_i occurs, N is the number of documents in a training set, and α is a smoothing constant.

The Naive Bayes model is called "Naive" because of its conditional independence assumption by assuming each word is conditionally independent of every other word given the category. Although this assumption is often violated in practice, the Naive Bayes model still tends to deliver competitive classification accuracy in a variety of applications compared to more sophisticated classifiers (Lewis, 1998).

- **Support Vector Machine (SVM)**

While being a relatively recent addition to a wide range of available classification techniques, the Support Vector Machine (Cortes & Vapnik, 1995) has been shown to be effective in the field of text mining (Y. Yang & Liu, 1999). SVMs are

classifiers that, in a binary classification problem, find a hyperplane in the instance space that has the maximum margin, i.e., the largest distance to the nearest training examples of any class (known as "functional margin"). For example, consider the three hyperplanes H_1, H_2, and H_3 in Figure 2.4; H_1 and H_2 are not good hyperplanes as H_1 does not properly separate the classes and H_2 separates the two classes only with a small margin. H_3 is so called "maximum-margin hyperplane" as it separates two classes with the maximum margin.

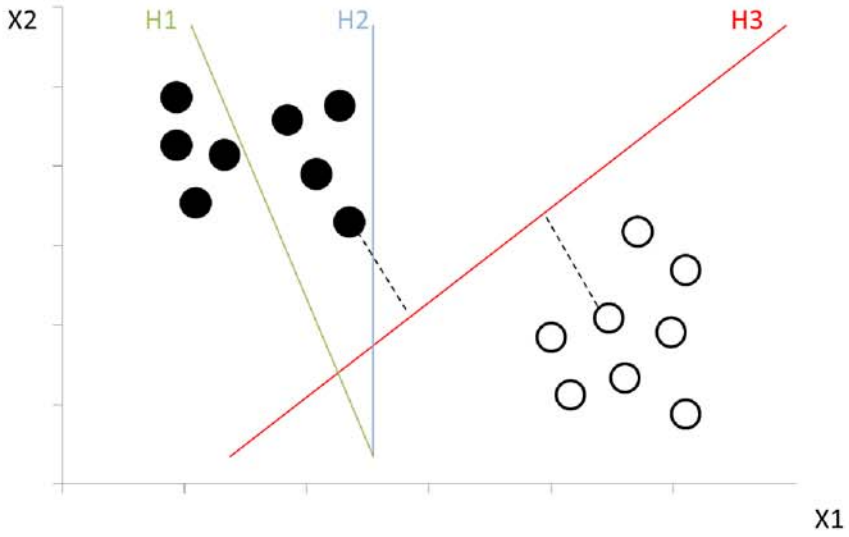

Figure 2.4: Explanation Example of SVM

- **Logistic Regression (LR)**

According to Mitchell (2016)'s textbook *Machine Learning*, Logistic Regression (LR) is a supervised learning approach to determine the function $f : X \Rightarrow Y$, or estimate the conditional probability $P(Y|X)$ where Y is a categorical dependent variable and $X = \langle X_1 \cdots X_n \rangle$ is any vector of real-valued features. This study focuses on the binary LR where Y is boolean.

By assuming that $P(Y|X)$ follows the form of logistic function, which outputs values that range between zero and one, the binary LR model can be expressed by

the following equations:

$$P(Y = 1|X) = \frac{1}{1 + exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \tag{2.10}$$

and

$$P(Y = 0|X) = \frac{exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \tag{2.11}$$

The value y_k that maximizes $P(Y = y_k|X)$ are assigned. That means that Y is assigned to 0 if the following condition holds:

$$1 < \frac{P(Y = 0|X)}{P(Y = 1|X)} \tag{2.12}$$

After substituting from previous two equations for $P(Y = 0|X)$ and $P(Y = 1|X)$ and taking the natural log of both sides, this equation leads to a simple linear equation for binary classification that assigns $Y = 0$ if X satisfies

$$0 < w_0 + \sum_{i=1}^{n} w_i X_i \tag{2.13}$$

and assign $Y = 1$ otherwise.

2.2 Challenges with High Dimensionality of Textual Data

2.2.1 Feature Selection Methods

Researchers have developed a variety of feature selection methods for dimensionality reduction. This section introduces four common feature selection methods associated in this study (Baharudin et al., 2010; Y. Yang & Pedersen, 1997).

1. Document Frequency Thresholding (DFT)

   Document Frequency Thresholding is applied by calculating the document frequency (DF) for every unique term that occurred in a training corpus and removing from the feature space the terms that has a DF lower than a given user-specified threshold. DF is the number of the documents in which a term occurs. While being the most commonly used and simplest technique for dimensionality reduction, the DFT is often considered as an ad-hoc approach because it is not theoretically grounded. In contrast to the Inverse Document Frequency (IDF) weight system (Sparck Jones, 1973), the basic assumption of DFT is that low-DF terms are not informative for predicting categories, which is contrary to a widely held belief in information retrieval that rare terms are more informative than common terms. It seems intuitively reasonable that common words deserve more consideration than rare ones. This can be traced back to Luhn's (1958) original idea. Luhn (1958) proposed a bell-shaped frequency distribution for most discriminatory words and suggested that middle-frequency words in a document are more indicative of its topicality, and that very common and very rare terms are weaker discriminators. Later other researchers also demonstrated that middle-ranked frequency words tend to be the best discriminators with the highest average document discriminatory power (Salton et al., 1975; van Rijsbergen, 1979). In real practice, many applications of statistical text analysis set upper and lower cut-off frequencies to remove these extremely high- and low-frequency words. In addition, excluding rare terms is theoretically effective

in reducing the feature space according to Zipf's law. Zipf (1949) claimed that "given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table" (Chennuru et. al., 2012).. Thus, a few words occur very often while many others occur rarely. By Zipf's law, removing the words that occur only in one single document (DF threshold = 1) would mean dropping half of the vocabulary and removing the words with the document frequency of less than 5 (DF threshold = 5) would mean that only 20% of the vocabulary is retained. Figure 2.5 shows the relationship between Zipf's Law and Luhn's Model.

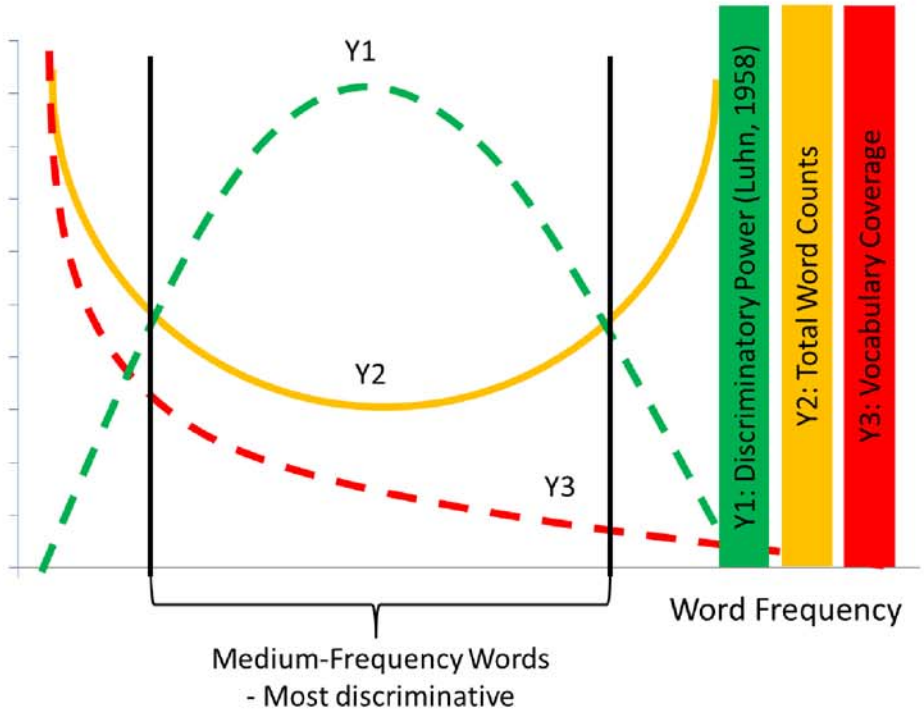

Figure 2.5: Luhns Bell-Shaped Model and Zipfs Law

Although most statistical text analysis methods follow the "standard" process of excluding low-frequency words in analyses and seem to work adequately well, the effectiveness of low-frequency words in statistical text anal-

ysis is still debatable as some empirical evidence has suggested that keeping rare terms can improve the performance of text clustering or classification. Price and Thelwall (2005) showed that removing the low-frequency words reduced the clustering power. Schnhofen and Benczr (2006) demonstrated that the extremely rare n-grams can be used to improve the classification accuracy. Al-Tahrawi (2013) found that the rare terms can enhance the text classification performance of polynomial networks classifier. However, these studies examined the impact of keeping rare terms in large corpora of well-written articles such as academic articles, news, or patents. None of the previous studies have examined the impact of keeping rare terms in text analysis for the noisy, unbalanced free-text data such as open-ended survey responses, where rare words are likely to be misspellings.

2. Odds Ratio (OR)

   Odds Ratio measures "the odds of the word occurring in the positive class c_i normalized by that of the negative class $\overline{c_i}$" (Simeon & Hilderman, 2008).

$$OR(t,c) = log\frac{odds(t|c_i)}{odds(t|\overline{c_i})} = log\frac{P(t|c_i)[1-P(t|\overline{c_i})]}{[1-P(t|c_i)]P(t|\overline{c_i})} \tag{2.14}$$

   where $P(c_i|t)$ is the probability of a category c_i given the presence of the word t and $P(c_i|\bar{t})$ is the probability of a category c_i given the absence of the word t.

   Some notations for the following two feature selection methods: Mutual information and Chi-square statistics. Considering a two-way contingency table of a term t and a category c in Table 2.2, A denotes the number of times a term t and a category c co-occur, B is the number of times a term t occurs without c, C is the number of times c occurs without t, D is the number of times c occurs without t, and N is the total number of documents.

Table 2.2: Two-way Contingency Table of Term and Category

N:Total # of documents	**Presence of t**	**Absence of t**
Membership in c	A	B
non-membership in c	B	D

3. Mutual Information (MI)

   Mutual Information (MI) is commonly used to measure the association between words using their co-occurrence information. In feature selection, MI can also be used to measure the association between a term t and a category c based on their statistical dependence, which is defined as follows:

$$I(t,c) = \log \frac{P(t \wedge c)}{P(t) \times P(c)} \approx \log \frac{A \times N}{(A+C) \times (A+B)} \tag{2.15}$$

   If a term t and a category c are independent, I(t,c) is zero. The goodness of a term in global feature selection can be measured through the average or maximum as follows:

$$I_{avg}(t) = \sum_{i=1}^{m} P(c_i) I(t,c_i) \tag{2.16}$$

$$I_{max}(t) = \max_{i=1}^{m} I(t,c_i) \tag{2.17}$$

   The score of MI can be significantly impacted by the marginal probabilities of terms and thus is not comparable across terms with a wide range of frequencies because MI favors low-frequency terms (Y. Yang & Pedersen, 1997).

4. Chi-square Statistics (CHI)

   The Chi-square statistics measure the lack of independence between a term t and a category c, which can be compared to the Chi-square distribution with one degree of freedom. The goodness of term t for a category c is defined as:

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{2.18}$$

   Similar to MI, the χ^2 statistic is close to zero if t and c are independent. Two alternative ways of measuring the goodness of a term t are:

$$\chi^2_{avg}(t) = \sum_{i=1} mP(c_i)\chi^2(t,c_i) \tag{2.19}$$

$$\chi^2_{max}(t) = \max_{i=1} m\{\chi^2(t,c_i)\} \tag{2.20}$$

   CHI is a normalized version of MI, thus $\chi^2(t,c)$ is comparable across terms for the same category. However, the χ^2 statistic is considered to be not reliable for low-frequency terms because of is proneness to overestimate the significance of relatively rare events. The reason comes from the fact that the assumption of normality does not hold for most real-world corpora unless enormous corpora are used or the analysis is restricted to only common words (Dunning, 1993).

2.2.2 Sparse Data Problem

In addition to the requirement of proper feature selection methods in statistical text analyses, another implication of the high-dimensionality of text is the proneness to the sparse data problem. Most statistical models have difficulty dealing with rarity especially as we have entered the era of big data. The "curse of dimensionality" is the phrase termed by Bellman (1957) in his mathematical optimization study, but this term has been widely used in many other domains to refer

to "various problems when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings" (Dirkmaat, 2013). Specifically, the volume of the feature space extremely increases with dimensionality, which results in a sparse feature space.

In statistical text mining, sparse features cause problems with identifying patterns in absolute and relative senses. In the absolute sense, these low-frequency features that happen to not occur in the training set may be mistakenly assigned a probability of zero. Oftentimes, these unseen events compose a great portion of a test set and thus negatively influence the classification performance. For instance, Essen and Steinbiss (1992) reported that in two text corpora with a 75/25 split, the fraction of word bigrams that occurred in the test set but did not occur in the training set is 12% for the million-word corpus and 50% for the other 100,000-word corpus. On the other hand, the problem with low-frequency features in relative sense is that even when rare features occur in the training set, due to their extremely low frequencies, any methods that require statistical significance may still fail to identify meaningful patterns between these rare features. Considering a rule association mining problem where two items that rarely occurs on their own often occur together when either one is present, the association between them may not be found because random co-occurrences are likely to swamp the meaningful associations between rare items (Liu et al., 1999). As a consequence, low-frequency features in sparse datasets inevitably increase the difficulty of identifying meaningful events (i.e., classes, cases, or patterns), especially the low-frequency ones.

The sparse data problem has negatively impacted the rare event identification. However, rare events "are often of great interest and great value" in practice (Weiss, 2004). There are two types of rare events that have been widely studied in the context of data or text mining:

- Rare classes are defined as the classes with a small coverage of cases. A data set is called unbalanced if at least one of the classes is represented by a signif-

icantly lower number of instances than others, which is known as the "class imbalance" problem (Ertekin, Giles, Storage, & Miscellaneous, 2007).

- Rare cases correspond to the instances that cover a meaningful but small region of the instance space different from the other members in the same class (Weiss, 2004).

Obviously, rarity has difficulty being identified due to its nature of scarcity, but the way most existing mining systems are designed also made it harder. Weiss (2004) identified a list of issues that increase the difficulty of identifying rare classes or rare cases from two perspectives: the nature of data (absolute / relative lack of data and noise) and the mining systems (improper evaluation metrics, inappropriate inductive bias, and data fragmentation).

1. Absolute rarity: absolute lack of data

   A fundamental problem with rarity is the shortage of data, which can arise from absolute and relative perspectives. Absolute rarity concerns the problems with the extremely small instance coverage or feature coverage in an absolute sense. Take rare cases for example, rare cases are more likely to be misclassified than common cases (Weiss, 1995). Because rare cases are more difficult to identify, most studies focus on their learned counterparts (i.e., small disjuncts) (Weiss, 2005). Empirical evidence suggests that rare cases often cause small disjuncts, which are found to have a much higher error rate than common disjuncts and collectively contribute a significant portion of classification errors in a test set (Holte, 1989; Weiss & Hirsh, 2000).

2. Relative rarity: relative lack of data

   The phrase "like a needle in a haystack" reflects the problem with the relative rarity. The difficulty of finding the needle is "not so much due to the needle being small or being only one needle but because the needle is obscured by the tremendous amount of hay" (Weiss, 2004). The rarely occurring instances

are usually overwhelmed by instances of the majority class so that they are much harder to identify. Considering a rule association mining problem, two items that rarely occurs on their own often occur together when either one is present. The association between them may not be found because random co-occurrences are likely to swamp the meaningful associations between rare items (Liu et al., 1999).

3. Noise

   The noise in data negatively influences the learning process of a classifier and tends to impact more on rare cases than on common cases. Due to the learner's ability to generalize, a rare case may fail to be learned if noisy data are mixed with the examples that are extremely few in the first place. A rare case may be learned if the learner is modified to generalize less, but noisy examples can also be mistakenly learned and covered by disjuncts. In this case, overfitting avoidance techniques such as cross-validation, regularization, or pruning are often involved to eliminate noise-induced small disjuncts and improve the learning process of "true" rare cases.

4. Improper evaluation metrics

   Evaluation metrics play an essential role in data or text mining. Metrics are used not only to evaluate the performance of a classifier but also to guide the learning process. Accuracy is a widely used evaluation metric but it fails to capture the poor classification performance of rare classes. For example, for a binary classification problem with a 90:10 class distribution, a classifier would give a seemingly satisfactory overall accuracy of 90% by simply classifying all examples into the majority class. In practice, precision and recall are considered better and fair metrics compared to accuracy. Precision is the fraction of all identified instances that truly belong to that class while recall is the fraction of the instances that belong to the class that are correctly identified. There is often a trade-off between precision and recall. F-measure (the

harmonic mean of precision and recall) is often recommended as it considers both metrics simultaneously. The minority class often has much lower precision and recall than the majority class. Many practitioners have observed the zero or near-zero recall for the minority class from extremely skewed class distributions (Choe, Lehto, Shin, & Choi, 2013).

5. Inappropriate inductive bias

   Rather than "memorizing" the examples, a good classifier can learn to generalize from the trend of examples to avoid overfitting. An overfitted model usually predicts very well on training data but often performs poorly on new or unseen data. However, proper generalization from a given training dataset often requires an extra-evidentiary bias. Many induction systems tend to favor common cases or classes due to their higher prior probabilities, which negatively impacts the ability to learn their rare counterparts. For example, typical decision tree learners tend to predict the most popular class and bias the results against rare classes in an unbalanced data set.

6. Data fragmentation

   Many data mining algorithms such as the decision tree classifier use a divide-and-conquer approach, where the original problem is recursively broken down into sub-problems. In such way, the instance space is partitioned recursively into smaller pieces, and thus regularities or patterns can be only found in these fragments of the instance space where only a small portion of data points are covered. This causes algorithms to require a large enough number of examples to give accurate probability estimates (Pagallo & Haussler, 1990); this problem gets worse when dealing with the identification of rare events.

3. RESEARCH OBJECTIVES AND METHOD

3.1 Research Objectives and Hypotheses

In text classification, the high-dimensionality of learning models has made the feature selection or dimensionality reduction inevitable. Prior to the modeling, two text preprocessing tasks have been widely implemented to reduce the feature space for training a classifier, including removing extremely high- and low-frequency words and applying a word normalization method such as stemming or lemmatization. However, the question of whether these features of extreme frequency should be removed is still debatable, as some studies have demonstrated better results by keeping these features while little study has systematically examined the impact. Thus, this study aims to fill the research gap by first exploring the role of high- and low-frequency features in text classification of injury narratives, and then proposing methods that address some of the limitations of stemming and lemmatization by grouping rare and unseen (words occurred only in test set) words linguistic features, morphologically and semantically.
In short, the objective of this study is two-fold:

(1) to explore the role of high- and low-frequency words in text classification of injury narratives, and

(2) to utilize rare and unseen words through grouping by their linguistic features (morphology and semantics) for better classification performance.

The first goal is to explore the role of low-frequency words (LFWs) in text classification. First of all, I explored the relationship between the document frequency of words and their importance for classification to determine whether the results were in conformity with the widely-held belief that LFWs should be excluded from

the analysis. The importance of words in text classification was determined by whether their absence negatively impacted the classification performance. A word was considered to be important for text classification if its absence (removal from feature space) decreased the classification performance. Thus, the importance of a word was measured by the impact of its removal, which demonstrated the classification performance difference between inclusion and exclusion of the word.

In addition, in the real-world application of machine learning, highly imbalanced data and limited availability of labeled data often hinder the classifiers performance. It is believed that low-frequency words are more valuable when they are present in smaller samples, for example, small categories or limited training data. Thus, it is also of the interest of this study to investigate the importance of low-frequency words in different levels of data availability in terms of the size of category and training/testing datasets.

In Chapter 4, I report the results of testing Hypotheses 1 to 4, regarding the importance of extreme-frequency features in text classification:

Null Hypothesis 1: *High-frequency words, specifically stopwords, are NOT important for text classification of injury data because the absence of them does not influence the classification performance.*
Alternative Hypothesis 1: *High-frequency words, specifically stopwords, are important for text classification of injury data because the absence of them can deteriorate the classification performance.*

Null Hypothesis 2: *Low-frequency words are NOT important for text classification of injury data because the absence of them does not influence the classification performance.*
Alternative Hypothesis 2: *Low-frequency words are important for text classification of injury data because the absence of them can deteriorate the classification performance.*

Null Hypothesis 3 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are important regardless of category sizes in text classification of injury data because the absence of them causes similar negative impact on the classification performance for any category sizes.*
Alternative Hypothesis 3 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are more important for small categories than large categories in text classification of injury data because the absence of them can cause more negative impact on the classification performance of small categories than large categories.*

Null Hypothesis 4 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are important for text classification regardless of sizes of training and test datasets, because the absence of them causes similar negative impact on the classification performance in any ratio of training data to test dataset sizes.*
Alternative Hypothesis 4 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are important for text classification when the training dataset size is limited rather than sufficient relative to the test dataset size, because the absence of them can cause more negative impact on the classification performance when the training dataset size is far smaller than the test dataset size, compared to when the training dataset size is far larger than the test dataset size.*

After exploring the importance of low-frequency words in text classification, I proposed methods to utilize low-frequency words for improving classification performance by considering their linguistic features. Classic word normalization methods such as stemming and lemmatization are commonly applied to group low-frequency words and reduce the size of feature space. These methods normalize words by removing their endings and mapping them to their stem or base form, assuming that words and their transformed forms (such as "electricity" and "electrical" and their stem "electr") carry similar semantic meaning and should predict the same category. Despite the popularity of stemming and lemmatization,

they have limitations to address misspellings, domain-specific words, or words without similar spellings. An ideal word normalization method should be able to group words that have same-hypernym (similar-concept) independent of their spelling. Conventional stemming and lemmatization only group same-hypernym words with similar spelling because, rather than considering semantics of words, they only consider their morphological similarity by assuming similar-spelling words mean similarly. In order to normalize or group words properly for statistical text analysis, knowing the semantic meaning of words is necessary. The literature has demonstrated the potential capability of statistical semantics in capturing the linguistic features of human natural language (Pecina, 2005, Han et al, 2011, Baroni et al., 2014); however, limited study has focused on further utilizing statistical semantics for text classification purpose. Thus, the second objective of this study is to investigate the practicality of statistical semantics in identifying same-hypernym words for grouping, with an ultimate goal of improving the performance of text classification with a smaller, denser, but more representative feature space.

Chapter 5 discusses the background and limitation of stemming and lemmatization in more detail and briefly introduces the proposed Type M+S Grouping Method. The following Chapters 6 and 7 are dedicated to the Type-M Morphological Mapping for grouping same-hypernym words with similar spelling and Type-S Semantic Grouping for grouping same-hypernym words with different spelling. In Chapters 7, I introduce two main branches of statistical semantic (i.e., correlational and distributional semantics) and then compare their effectiveness in identifying same-hypernym words. An associated hypothesis is listed in Hypothesis 5:

Null Hypothesis 5: *There is no difference between the effectiveness of distributional semantics and correlational semantics in identifying same-hypernym words.*
Alternative Hypothesis 5: *Distributional semantics is more effective than correla-*

tional semantics in identifying same-hypernym words.

In Chapter 8, I report a test of three strategies that can potentially improve the proposed Type M+S Grouping Method and to compare the effectiveness of the proposed methods and classic word normalization methods (stemming and lemmatization) in improving the text classification performance. The results can be used to prove or disprove Hypotheses 6 and 7 in the following:

Null Hypothesis 6: *There is no difference between the proposed Type M+S Grouping Method and traditional word normalization methods, stemming and lemmatization, in improving classification performance of injury data.*
Alternative Hypothesis 6: *The proposed Type M+S Grouping Method is superior to traditional word normalization methods, stemming and lemmatization, in improving classification performance of injury data.*

Null Hypothesis 7: *The Proposed Type M+S Grouping Method cannot be further improved by the different add-on features, including:*

- Considering sequences of words
- Manual review
- Applying Naive Bayes-weighted Support Vector Machine (SVM) and Logistic Regression (LR) Classifiers

Alternative Hypothesis 7: *The Proposed Type M+S Grouping Method can be further improved by the different add-on features, including:*

- Considering sequences of words
- Manual review
- Applying Naive Bayes-weighted Support Vector Machine (SVM) and Logistic Regression (LR) Classifiers

3.2 Data

The dataset used in this study was the injury narratives collected from 2002 to 2012 by the Queensland Injury Surveillance Unit, located in Queensland, Australia. In this so-called QISU dataset, more than half a million injury narratives along with several injury related variables (manually-labeled external cause and mechanism of injury) were recorded by the nurses from emergency departments in public hospitals across Queensland.

The categories to predict were modified based on the External Cause of Injury provided along with the injury narratives. Table 3.1 lists all categories for prediction and their distribution, the original external cause code, and description. Some external causes were combined due to their similar nature while some comprehensive external causes (18 for other and unspecified types of poisoning and allergy and 28 for other injuries) were further classified into sub-categories for distinguishable differences.

The injury cases due to the causes other than drug or medical substances (External Cause Code 17 or "PA_DRUG") were originally coded 18 for External Cause Code. These causes were further classified into 6 sub-categories according to the different types of injury agents, including: alcohol, chemical, drug plus chemical, food, plant, and others. With a name starting "PA", which stands for poisoning and allergy, the corresponding new categories were annotated with: "PA_ALCOHOL", "PA_CHEMICAL", "PA_DRUGALCO", "PA_FOOD", "PA_PLANT", and "PA_OTHERS."

In addition, the injury cases other than these main causes listed in Table 3.1 were coded 28 and 29 for External Cause Code. A large portion of these cases were related to foreign body, and thus further classified into sub-categories depending on the body part contacted (eye, ear, nose, mouth). The four new foreign body related categories were annotated with: "C289EYE" for foreign body in eye and eye related injury, "C289FBEAR" for foreign body in ear, "C289FBI" for foreign

body ingested in upper digestive track including mouth, esophagus, and stomach, and "C289FBNOSE" for foreign body in nose.

Figure 3.1 visually displays the distribution of the categories for prediction. These categories were grouped into three category sizes in order to study their differences among different experimental settings. The size of a category is considered *small* if the percentage of relevant cases relative to all cases in the QISU dataset is less than 0.33%, *medium* if between 0.33% and 10%, and *large* if greater than 10%. The QISU dataset is a typical imbalanced dataset, with a few dominating large categories and many small categories.

Table 3.1: External Cause Category List, Distribution, and Description

Category of External Causes	# of Documents	% of Total Cases	Original External Cause Code	External Cause Description
ANIMAL	20,033	4.01%	21 and 22	Dog related and animal related excluding horse
BICYCLE	13,138	2.63%	5	Pedal cyclist or pedal cycle passenger
C289	46,201	9.24%	28 and 29	Other and unspecified external cause
C289EYE	10,103	2.02%	28	Foreign body in eye and eye related injury
C289FBEAR	2,098	0.42%	28	Foreign body in ear
C289FBI	3,769	0.75%	28	Foreign body ingested
C289FBNOSE	2,553	0.51%	28	Foreign body in nose
CHOKING	1,157	0.23%	13	Other threat to breathing
CUTTING	35,498	7.10%	20	Cutting, piercing object
DROWNING	450	0.09%	11 and 12	Drowning, submersion
ELECTRICITY	1,312	0.26%	25	Electricity
FALL	158,577	31.71%	9 and 10	Fall- low and high
FIREARM	72	0.01%	19	Firearm
FIREFLAME	1,596	0.32%	14	Fire, flames, smoke
HORSE	3,785	0.76%	8	Horse Related (fall from, struck or bitten by)
HOTCOLDCOND	461	0.09%	26 and 27	Hot and cold conditions
HOTOBJ	9,658	1.93%	15 and 16	Exposure to hot drink, food, water, other fluid, steam, gas, vapour, object, solid substance
MACHINERY	17,511	3.50%	24	Machinery
MOTORCYCLE	11,841	2.37%	3 and 4	Motorcycle driver and passenger
MOTORVEHICLE	14,940	2.99%	1 and 2	Motor vehicle – driver and passenger
OTHERTRANSPORT	2,786	0.56%	7	Other or unspecified transport related circumstance
PA_ALCOHOL	543	0.11%	18	Poisoning and allergy due to alcohol
PA_CHEMICAL	2,858	0.57%	18	Poisoning and allergy due to chemicals
PA_DRUG	5,578	1.12%	17	Poisoning and allergy due to drug or medicinal substance
PA_DRUGALCO	943	0.19%	17 and 18	Poisoning and allergy due to drug and alcohol
PA_FOOD	870	0.17%	18	Poisoning and allergy due to food
PA_OTHERS	1,504	0.30%	18	Poisoning and allergy due to other or unspecified substance
PA_PLANT	220	0.04%	18	Poisoning and allergy due to grass, tree, bush, plant
PEDESTRIAN	1,920	0.38%	6	Pedestrian
STRUCKCOLLISION	128,142	25.62%	30 and 31	Struck by or collision with person or object

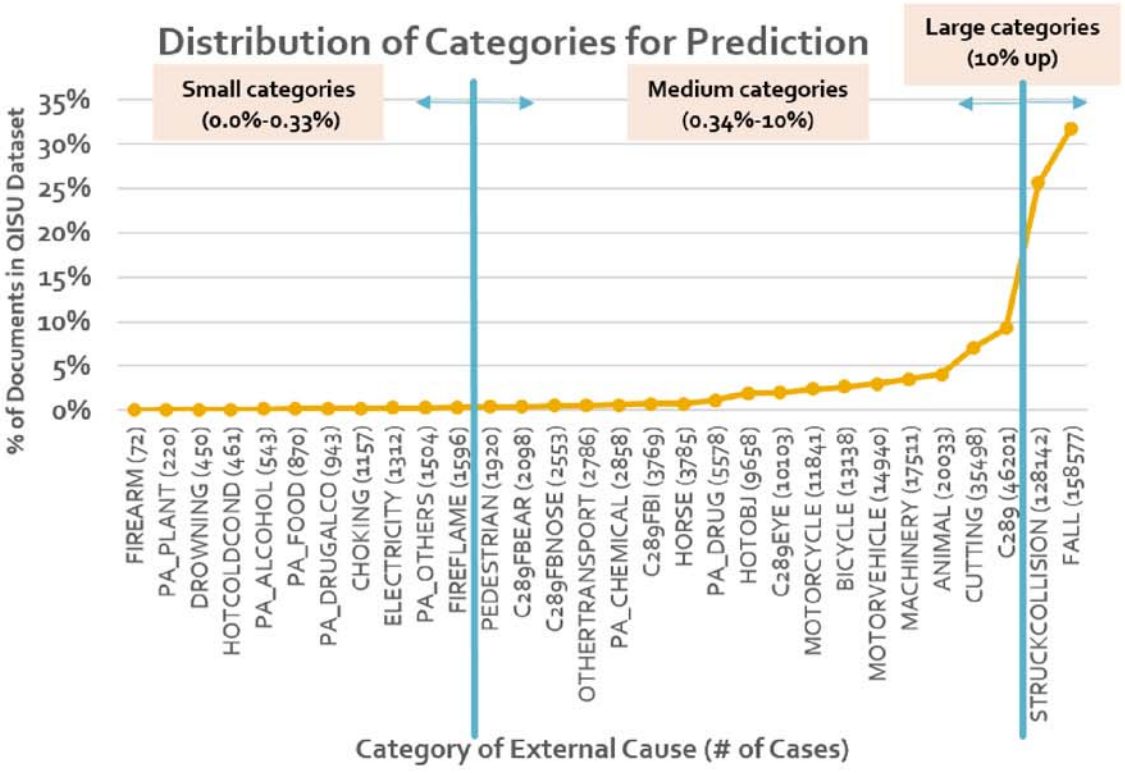

Figure 3.1: External Cause Category Distribution of QISU Dataset

Table 3.2: Statistics Summary of QISU Dataset

	# of documents	# of word occurrences	# of unique words (vocabulary size)	Avg. # of words per document	# of classes
QISU	500,117	6,536,611	47,420	13	30

This QISU injury dataset is a half-a-million collection of injury narratives with 6.5-million word occurrences and 47-thousand unique words. Each injury narrative has 13 words on average (range: 1 to 51 words). The half-million injury narratives were coded into one of 30 categories of external causes. The statistics for the QISU dataset are summarized in Table 3.2.

The vocabulary size (i.e., the number of unique words) of the QISU dataset is 47,420. By counting the document frequency (DF) (i.e., the number of documents each word in the vocabulary set occurs) and how many words have this DF, the results supported the proposition of Zipf (1949)'s law that the proportion of vocabulary made up by words with a certain frequency f should equal the inverse of $f(f+1)$. Figure 3.2 below shows the relationship between words' DF and actual / predicted vocabulary coverage for the top 20 lowest DF. The numeric information of vocabulary coverage is listed in Table 3.3, which shows the consistent pattern between the actual and predicted.

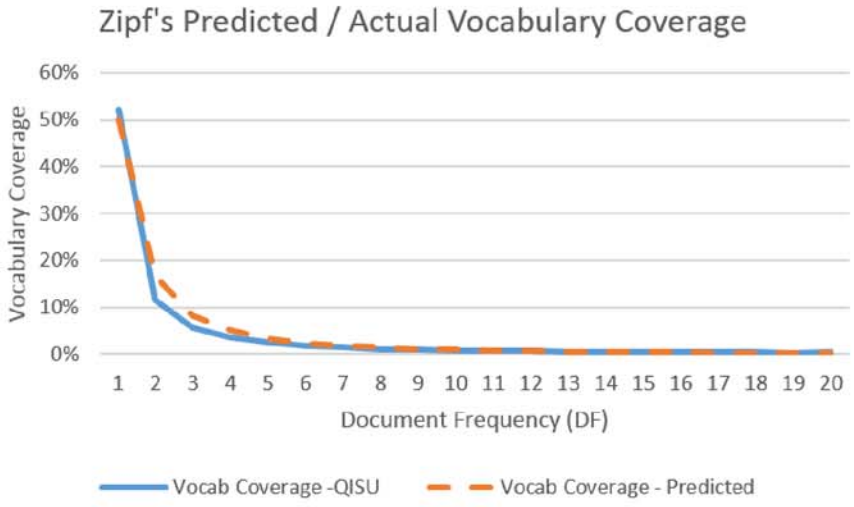

Figure 3.2: Relationship between Word Frequency and Vocabulary Coverage: Predicted vs. Actual

Table 3.3: Actual vs. Predicted Vocabulary Coverage for DF of 1 to 20

DF	# of words	Vocab. Coverage-Actual	Vocab. Coverage- Predicted
1	24683	52.05%	50.00%
2	5483	11.56%	16.67%
3	2687	5.67%	8.33%
4	1646	3.47%	5.00%
5	1214	2.56%	3.33%
6	862	1.82%	2.38%
7	725	1.53%	1.79%
8	527	1.11%	1.39%
9	415	0.88%	1.11%
10	389	0.82%	0.91%
11	351	0.74%	0.76%
12	315	0.66%	0.64%
13	260	0.55%	0.55%
14	241	0.51%	0.48%
15	223	0.47%	0.42%
16	197	0.42%	0.37%
17	186	0.39%	0.33%
18	202	0.43%	0.29%
19	155	0.33%	0.26%
20	171	0.36%	0.24%

3.3 Method

This study focused on the supervised learning, specifically the text classification of injury narratives. The injury narratives of the QISU dataset were manually coded into one of thirty categories of external cause. These category labels served as the "gold standard" to compare with the machine-predicted answers.

Several methods were proposed and tested through the study. Their effectiveness were measured by the difference in classification performance compared to standard classification that did not involve any text preprocessing tasks other than lowercase transformation and removal of non-English characters.

To test the proposed method in improving the classification of injury narrative, three classic supervised classifiers were used for training and prediction: Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), and Logistic Regression (LR). These three classifiers have been widely implemented in a variety of classification tasks including injury data (Bertke et al., 2016; Nanda et al., 2016; Vallmuur, 2015). The theoretical basis of these classifiers are briefly defined in Section 2.1.5. Various software packages for machine learning are publically available. With these off-the-shelf packages, a machine learning classifier can be trained to learn from labeled training data, analyze, and make predictions on the unlabeled test set. This study used the Python software along with its natural language processing and machine learning related packages because they are free to the public and easy to implement with flexibility. The three classifiers used in this study were built with Pythons Scikit-learn package (Pedregosa, 2011), by using the default setting of the $sklearn.naive\_bayes.MultinomialNB$ module for MNB, $sklearn.svm.LinearSVC$ module for SVM, and $sklearn.linear\_model.LogisticRegression$ module for LR. Other Python packages used in this study included NLTK (Bird, Loper, & Klein, 2009) for natural language processing and Gensim (Rehurek & Sojka, 2010) for developing models of Word2Vec (Mikolov et al., 2013).

To evaluate the effectiveness of proposed methods at different accessibility levels of training and test data, this study measured the classification performance of proposed methods in three data scenarios represented by the train-test ratio (i.e., the ratio of training cases to test cases). Three train-test ratios for testing were: 1:9 (training data are limited as little as one-tenth of the test data), 1:1 (training data and test data are comparable in quantity), and 9:1 (training data are as much as

nine times of test set). Table 3.4 lists the corresponding numbers of train and test cases for each train-test ratio scenario.

Table 3.4: Numbers of Training and Test Cases in Train-Test Ratios of 1:9, 1:1, 9:1

Category (Category Size)	# Total Cases	% Corpus	1:9		1:1		9:1	
			# Train	# Test	# Train	# Test	# Train	# Test
FIREARM (S)	72	0.01%	65	7	36	36	7	65
PA_PLANT (S)	220	0.04%	198	22	110	110	22	198
DROWNING (S)	450	0.09%	405	45	225	225	45	405
HOTCOLDCOND (S)	461	0.09%	415	46	231	230	46	415
PA_ALCOHOL (S)	543	0.11%	489	54	272	271	54	489
PA_FOOD (S)	870	0.17%	783	87	435	435	87	783
PA_DRUGALCO (S)	943	0.19%	849	94	472	471	94	849
CHOKING (S)	1,157	0.23%	1,041	116	579	578	116	1,041
ELECTRICITY (S)	1,312	0.26%	1,181	131	656	656	131	1,181
PA_OTHERS (S)	1,504	0.30%	1,354	150	752	752	150	1,354
FIREFLAME (S)	1,596	0.32%	1,436	160	798	798	160	1,436
PEDESTRIAN (M)	1,920	0.38%	1,728	192	960	960	192	1,728
C289FBEAR (M)	2,098	0.42%	1,888	210	1,049	1,049	210	1,888
C289FBNOSE (M)	2,553	0.51%	2,298	255	1,277	1,276	255	2,298
OTHERTRANSPORT (M)	2,786	0.56%	2,507	279	1,393	1,393	279	2,507
PA_CHEMICAL (M)	2,858	0.57%	2,572	286	1,429	1,429	286	2,572
C289FBI (M)	3,769	0.75%	3,392	377	1,885	1,884	377	3,392
HORSE (M)	3,785	0.76%	3,406	379	1,893	1,892	379	3,406
PA_DRUG (M)	5,578	1.12%	5,020	558	2,789	2,789	558	5,020
HOTOBJ (M)	9,658	1.93%	8,692	966	4,829	4,829	966	8,692
C289EYE (M)	10,103	2.02%	9,093	1,010	5,052	5,051	1,010	9,093
MOTORCYCLE (M)	11,841	2.37%	10,657	1,184	5,921	5,920	1,184	10,657
BICYCLE (M)	13,138	2.63%	11,824	1,314	6,569	6,569	1,314	11,824
MOTORVEHICLE (M)	14,940	2.99%	13,446	1,494	7,470	7,470	1,494	13,446
MACHINERY (M)	17,511	3.50%	15,760	1,751	8,756	8,755	1,751	15,760
ANIMAL (M)	20,033	4.01%	18,030	2,003	10,017	10,016	2,003	18,030
CUTTING (M)	35,498	7.10%	31,948	3,550	17,749	17,749	3,550	31,948
C289 (M)	46,201	9.24%	41,581	4,620	23,101	23,100	4,620	41,581
STRUCKCOLLISION (L)	128,142	25.62%	115,328	12,814	64,071	64,071	12,814	115,328
FALL (L)	158,577	31.71%	142,719	15,858	79,289	79,288	15,858	142,719

3.4 Performance Evaluation

To evaluate the classification performance, this study used F-measure, which is a typical measure in information retrieval. F-measure, also known as F1-score, is a mixed measure that considers both precision and recall. Precision is the fraction of all labeled comments that truly belong to that category while recall is the fraction of the documents that belong to the category that are correctly labeled. There is often a trade-off between precision and recall. This study intended to consider both measures at the same time. As a result, F-measure, the harmonic mean of precision and recall, was used to measure classification performance for each category for prediction.

Precision and recall are calculated in the following equations:

$$precision = \frac{CL}{TL}$$

$$recall = \frac{CL}{TC}$$

where CL (Correct Label) is the number of documents labeled with the correct category, TL (Total Label) is the number of documents claimed to be in the category, and TC (Total Category) is the number of documents that actually belong to the category.

Given precision and recall, F-measure can be calculated by the equation below:

$$F-measure = 2 \times \frac{precision \times recall}{precision + recall}$$

With the F-measure to evaluate the classification performance of each individual category, the overall classification performance of all categories was evaluated based on the macro-averaged F-measure across categories in this study. The macro-averaged F-measure is an unweighted average of F-measures of all targeted categories. This way, each category has equal weight towards the overall results regardless of size and thus, the bias resulting from large categories dominating the performance results can be avoided. The macro-averaged F-measure across a total of C categories is:

$$Macro-averaged\ \ F-measure = \frac{\sum_i^C F_i}{C}$$

where F_i is the F-measure of category i for $i = 1,\ldots,C$ and C is the number of categories for prediction.

4. ROLE OF EXTREME-FREQUENCY WORDS IN TEXT CLASSIFICATION OF INJURY NARRATIVES

In natural language processing, the Vector Space Model (VSM) is used to structure a text corpus so that a computer can process and train a learning classifier. Textual data are presented, from unstructured to structured, by a word-by-document matrix that records the frequency of each word occurring in each document in the corpus. The feature space is often built upon the entire vocabulary (i.e., unique words) of a corpus. A corpus that has tens of thousands of features is common in practice. The resulting problems related to the huge but sparse VSM, the so-called "curse of dimensionality" coined by Bellman (1957), has imposed many limitations on the performance of machine learning and therefore, has often requires the implementation of feature selection in text preprocessing to reduce dimensionality.

In 1958, Luhn proposed a bell-shaped frequency distribution model for most discriminatory words, claiming that extreme-frequency words have the least discriminatory power in statistical text analysis. Extremely high- and low-frequency features contribute greatly to the sparsity of the VSM. Zipf (1949) also claimed that, given a corpus of natural language utterances, a few words occur frequently while many others occur rarely, which was consistent with the findings in Chapter 3. Those extremely high-frequency features (specifically, stopwords or functional words such as grammatical articles or prepositions) take up a great portion of word occurrences in the corpus whereas extremely low-frequency features comprise the majority of the feature space (words that occurred only once contributed to half of the vocabulary). Thus, removing these extreme-frequency features has become a standard text preprocessing task that is commonly implemented prior to training a classifier. Some studies, however, have demonstrated that these extreme features

were indeed important for statistical text analysis and keeping them resulted in better performance. Due to the lack of systematical study in this field, one objective of this study is to provide empirical evidence by investigating the importance of extremely high- and low-frequency words in text classification. As mentioned in Chapter 3, the importance of words is evaluated by the impact on classification performance due to their absence from the feature space.

The following sections explore the importance of extreme-frequency words in classifying injury narratives and report the testing results of Hypotheses 1 to 4, with Section 4.1 for high-frequency words and Section 4.2 for low-frequency words.

4.1 High-frequency Words

4.1.1 Background

Some high-frequency words that carry semantic meaning are actually critical to text analysis. For example, the word "fall" has a very high frequency in the QISU dataset and is an indicative feature of fall-related injuries. A machine learning classifier is less likely to properly learn and classify fall-related injury cases without this discriminatory feature "fell." However, many other high-frequency words are stopwords, which are function words that have little semantic meaning but essential to maintain the grammatical relationship with other words. Stopwords often contribute to a large portion of high frequency words and also total word occurrences of a corpus. Thus, stopwords are often used interchangeably with high-frequency or common words, although they are not equivalent because the high-frequency words are defined based on word frequency in a specific collection. Given that some high-frequency words that carry semantic meaning are indeed discriminatory for classification, this study limits the high-frequency words to stopwords, which do not carry semantic meaning and have questionable discriminatory power for classifying injury narratives.

In practice, stopwords are often filtered out and excluded from model development due to the lack of meaning and high occurrence in text data (Manning et al., 2008). Typical stopwords include prepositions, pronouns, auxiliary verbs, conjunctions, and grammatical articles. Various English stopword lists are available online to the public, including: "SMART stopword list" built by Gerard Salton and Chris Buckley for the experimental SMART information retrieval system at Cornell University, "Snowball stopword list" from the Snowball stemming project, and "NLTK (Natural Language Toolkit) stopword list" from the NLTK corpus implemented in Python. In addition to using an off-the-shelf stopword list, some machine learning practitioners exclude stopwords by setting a frequency cutoff and remove words with a frequency higher than the threshold, or create their own stopword list by sorting words by frequency in the document collection and filtering out the most frequent ones that do not carry meaning related to the domain of the documents being analyzed.

While removing stopwords seems to become a standard text preprocessing task, some studies have showed that stopwords are indeed critical for text classification and do contribute to the meaning of phrases. Riloff (1995) argued that the prepositions and auxiliary verbs play an important role in building complex phrases essential to text classification of the joint-ventures domain. The presence of the preposition provides a more "specific" concept while the pairing words could not do so by themselves. For example, the presence of the preposition "with" along with the words "venture" and "joint venture" almost always implies the involvement with a specific partner while the two words by themselves do not necessarily represent a specific joint venture activity. Another study from McCallum and Nigam (1998) showed that the pronoun "my" is a discriminative feature to identify student home pages. In addition, sentiment analysis often uses the stopwords relating to negation (e.g. *no*, *not*, *don't*, *can't*, etc.) to identify the sentiment of a sentence and removing them can negatively impact the performance (Pak & Paroubek, 2010; Saif et al., 2014).

4.1.2 Experimental Design and Results

As there has been little study on the effect of removing stopword on the classification of injury surveillance data, this study aims to fill the research gap by examining the role of stopwords in the injury classification domain.

As the first step of study, I selected and categorized stopwords from the common stopword lists into five types based on their grammatical roles, as shown in Table 4.1.

Table 4.1: Types and Lists of Frequently-Occurring Stopwords

Stopword Type	Stopwords
Grammatical articles	*a, an, the,*
Pronouns	*i, me, my, myself, we, us, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those*
Auxiliary verbs	*am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, will, would, shall, should, can, could, may, might, must, ought, im, youre, hes, shes, its, were, theyre, ive, youve, weve, theyve, id, youd, hed, shed, wed, theyd, ill, youll, hell, shell, well, theyll, ve, re, ll, isnt, arent, wasnt, werent, hasnt, havent, hadnt, doesnt, dont, didnt, wont, wouldnt, shant, shouldnt, cant, cannot, couldnt, mustnt, isn, aren, wasn, weren, hasn, haven, hadn, doesn, don, didn, won, wouldn, shan, shouldn, can, cannot, couldn, mustn*
Prepositions	*of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, onto, into, across*
Others	*and, but, if, or, because, as, until, while, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, again, further, then, once, left, right*

The importance of identified stopword types was determined by the impact on classification performance due to its absence. The impact of removing stopwords

on text classification was then evaluated based on the difference between removing and keeping stopwords in an average of 10-fold cross-validation classification performance. The classification performance was measured by the macro-averaged F-measure of all categories, classified by three classic classifiers (MNB, SVM, LR) in three train-test split scenarios (1:9, 1:1, 9:1).

The main factor, as an independent variable annotated with "RemovedStopwordType," was the type of stopwords being removed in the experiment:

- Grammatical articles (article)
- Pronouns (pronoun)
- Auxiliary verbs (aux. verb)
- Prepositions (preposition)
- Others such as common adverbs and indefinite pronouns (other)

Other secondary factors (and their potential interactions) that could potentially influence the classification performance were also considered, which were: $Classifier$ (MNB, SVM, LR), $Train-Test\_Ratio$ (1:9, 1:1, 9:1), and $CategorySize$ (L, M, S). The dependent variable was the impact, measured by the difference in the macro-averaged F-measure.

The analysis of variance (ANOVA) test was conducted to determine if there were any statistically significant differences between the means of F-measure in the cases of not removing stopwords and five stopword removal scenarios (removing stopwords of article, pronoun, auxiliary verb, preposition, and other). The results of the ANOVA test are listed in Table A.1 in Appendix A, showing that removing any types of stopwords did not statistically significantly impact the classification performance (F-value = 0.15, P-value = 0.9798).

Another ANOVA test was conducted to determine if there were any statistically significant differences between the impact means among different stopword

types $RemovedStopwordType$. The results are listed in Table A.2 in Appendix A, suggesting that removing different types of stopwords had a statistically different impact on F-measure (F-value = 5.56, P-value = 0.0002).

The ANOVA results also indicated that the three-way $RemovedStopwordType * Classifier * Category\_Size$ interaction was significant at an alpha level of 0.05 (F-value = 10.61, P-value <0.0001). The interaction between the three factors was examined visually with three graphs of two-way interaction between Category_Size and RemovedStopwordType at each level of Classifier (MNB, SVM, LR). Figures 4.1-4.3 are the two-way interaction plots, one for each classifier: MNB, SVM, and LR.

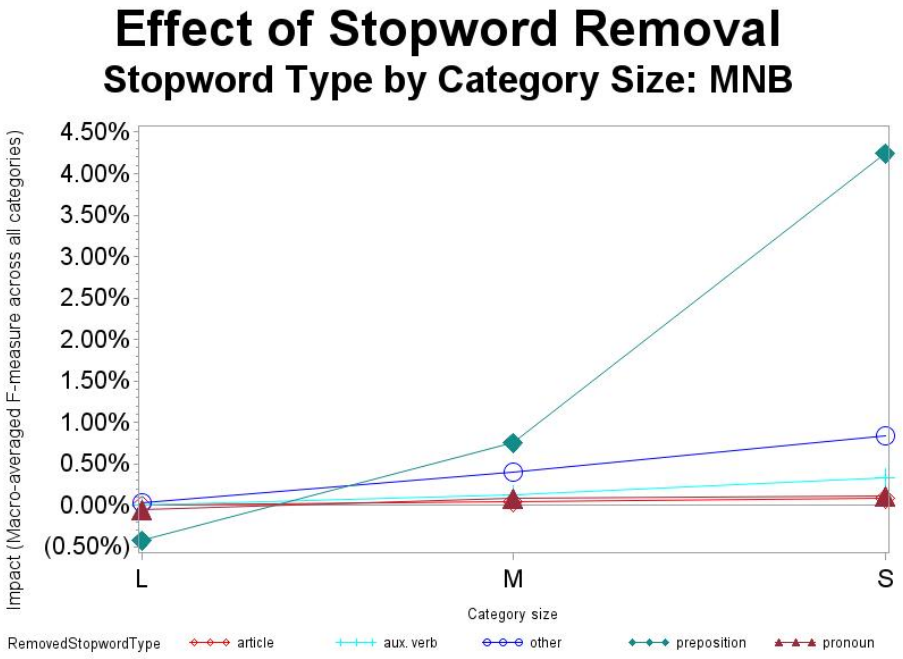

Figure 4.1: Effect of Stopword Removal on Classification of MNB by Category Size

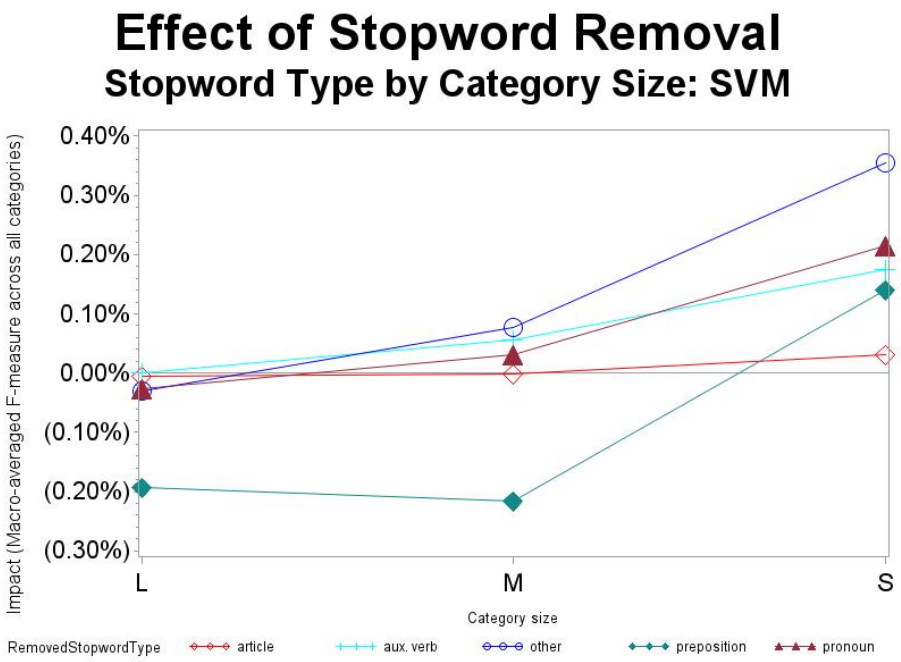

Figure 4.2: Effect of Stopword Removal on Classification of SVM by Category Size

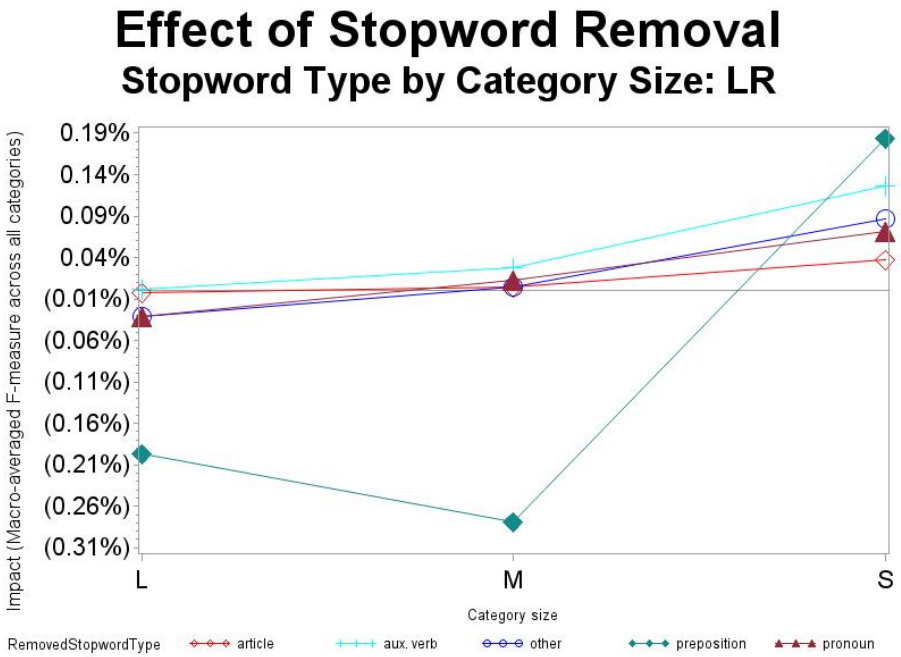

Figure 4.3: Effect of Stopword Removal on Classification of LR by Category Size

Overall, removing stopwords tended to have a positive impact on small categories, slight positive on medium categories, and insignificantly negative on large categories. Among five types of stopwords, removing preposition stopwords appeared to be the most influential in terms of causing the highest negative impact on large categories while the highest positive impact on small categories for MNB and LR. The only exception was found in the small categories classified by SVM where removing preposition stopwords did not show the highest improvement compared to removing other types of stopwords.

The ANOVA results indicated that there were significant statistical differences among 5 types of stopwords. Thus, two post hoc tests (Tukey and LSD) were conducted for multiple comparison. As Table A.3 shows, both tests had similar grouping results in which removing preposition stopwords achieved the greatest positive improvement. However, it was observed from Figures 4.1-4.3 that removing preposition stopwords also caused a negative impact on the classification performance of medium and large categories for SVM and LR. Figure 4.4 shows the distribution of impact on classification performance due to the stopword removal, which, again, suggests the greatest influence, both positive and negative, that removing preposition stopwords can cause.

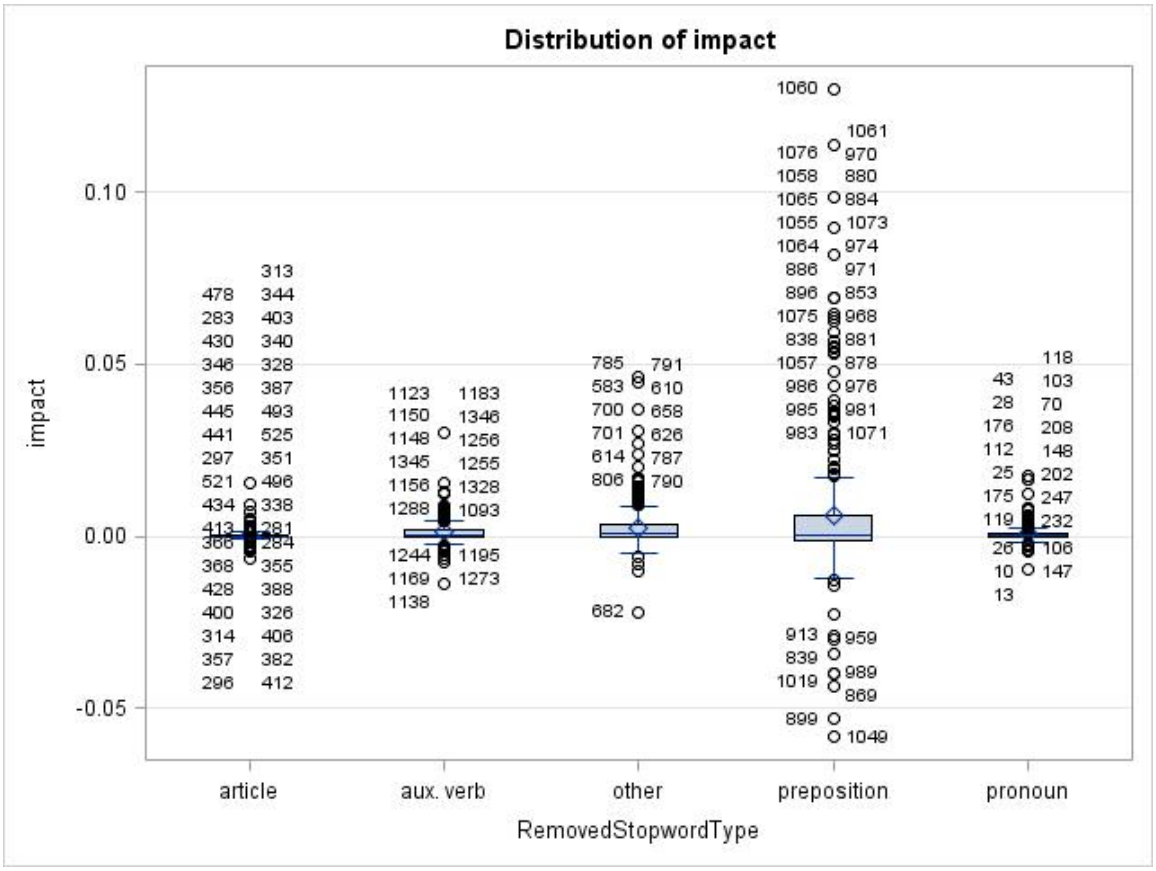

Figure 4.4: Distribution of Impact Caused by Stopword Removal

By investigating the data points on both ends of long tails, removing preposition stopwords greatly improved the classification of many small categories (e.g. DROWNING, ELECTRICITY, PA_FOOD, CHOKING) for MNB; however, it significantly negatively influenced the category PEDESTRIAN, consistently for all three learning classifiers in all three test-train scenarios. Table 4.2 tabulates the resulting negative impact on classification performance of the PEDESTRIAN category for three classifiers in three train-test ratio scenarios.

Table 4.2: Impact of Removing Preposition Stopwords on F-measure of PEDESTRIAN Category

Impact	**Train-Test Ratio**			**Overall**
Classifier	**1:9**	**1:1**	**9:1**	
MNB	-0.7%	-4.0%	-5.3%	-3.3%
SVM	-4.0%	-2.3%	-3.4%	-3.2%
LR	-5.8%	-3.0%	-4.4%	-4.4%

The reason for failing to properly classify cases of the PEDESTRIAN category after removing proposition stopwords was because a great amount of PEDESTRIAN cases involved expressions such as "run over by car", "hit by car", and "struck by car". Without the preposition "by", these PEDESTRIAN cases were often predicted as MOTORVEHICLE cases by classifiers. The presence of the word "by" along with "car" implies the involvement of something else (usually the injured person in PEDESTRIAN cases) interacting with a car. Because the word "car" served as a discriminative feature for identifying the MOTORVEHICLE category, a PEDESTRIAN case with the word "car" occurring alone without "by" was likely to be misclassified as MOTORVEHICLE, which was a much bigger category that most classifiers were biased in favor of. The results here were consistent with Riloff (1995)s finding regarding identifying joint venture activity: the prepositions were essentially important in building complex phrases to provide specific concept that was often different from the pairing word alone.

Last but not least, Hypothesis 1, listed below, was tested given the results above.

Null Hypothesis 1: *High-frequency words, specifically stopwords, are NOT important for text classification of injury data because the absence of them does not influence the classification performance.*

Alternative Hypothesis 1: *High-frequency words, specifically stopwords, are impor-*

tant for text classification of injury data because the absence of them can deteriorate the classification performance.

Null Hypothesis 1 was falsified and Alternative Hypothesis 1 was accepted because the absence of preposition stopwords significantly decreased the classification performance of the PEDESTRIAN category, consistently for three classifiers in three train-test ratio scenarios, although removing preposition stopwords greatly improved many small categories (which were performed badly originally) for MNB and showed positive overall impact.

4.2 Low-Frequency Words

4.2.1 Background

In this section, the importance of low-frequency words (LFWs) in text classification is explored. Many machine learning applications exclude LFWs as their feature selection method to reduce the dimensionality of feature space. However, some studies have demonstrated an improvement in statistical text analysis by keeping LFWs (Al-Tahrawi, 2013; Price & Thelwall, 2005; Schnhofen & Benczr, 2006). There has been little research on the role of LFWs in text classification, not to mention in the field of safety and injury. Thus, this research aims to provide empirical evidence in order to answer the fundamental question: whether to keep or remove LFWs in text classification of injury data?

In practice, the majority of datasets are imbalanced with many small categories and few large categories. In addition, the availability of labeled training data is often limited compared to ever-growing volumes of unlabeled data. Most free-text data are not structured or coded in a way corresponding to actual need for analysis. Therefore, this study also examined the relationship between LFWs and scarcity of data in text classification. In other words, this study explored the effect of removing LFWs from the perspectives of relative sizes of categories (*Category Size*)

and relative sizes between training and test sets ($Train-test\ Ratio$). Considering the nature of rarity, it is reasonable to think that LFWs should be more valuable for small categories and more important when the training dataset is limited in quantity relative to the prediction dataset.

4.2.2 Experimental Design and Results

One objective of this study is to explore the importance of low-frequency words (LFWs) in text classification. As noted in Chapter 3, the importance of a word was evaluated by the impact on the classification performance due to its absence.

To investigate the impact of removing LFWs on classification performance, the macro-averaged F-measure of all categories was measured at different levels of document frequency cut-off (DFC), for three classic classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). The range of DFC in this investigation was from 1 to 40. Words with a document frequency (DF) lower than a pre-selected DFC were excluded from classifier development. When DFC = 1, no words were removed. When DFC = 2, words that occurred only in one document, called "DF1 words," were removed, and so on. Words that had a DF of x (i.e., occurred in x documents) were denoted with "DF x words" for simplicity in this study. For example: "DF1 words" were the words that occurred only in one document, which were removed when DFC = 2.

Figure 4.5 shows the overall classification performance for 3 classifiers (MNB, SVM, LR), averaged over three train-test ratios, at pre-set levels of DFC from 1 to 40. The figure displays two patterns of the effect of removing LFWs. For LR and SVM, the performance gradually decreased as DFC increased (i.e., LFWs were continually removed). For MNB, on the other hand, the performance dramatically increased first at the DFC2 level, then increased at a declining rate until DFC15, and finally started decreasing gradually afterwards.

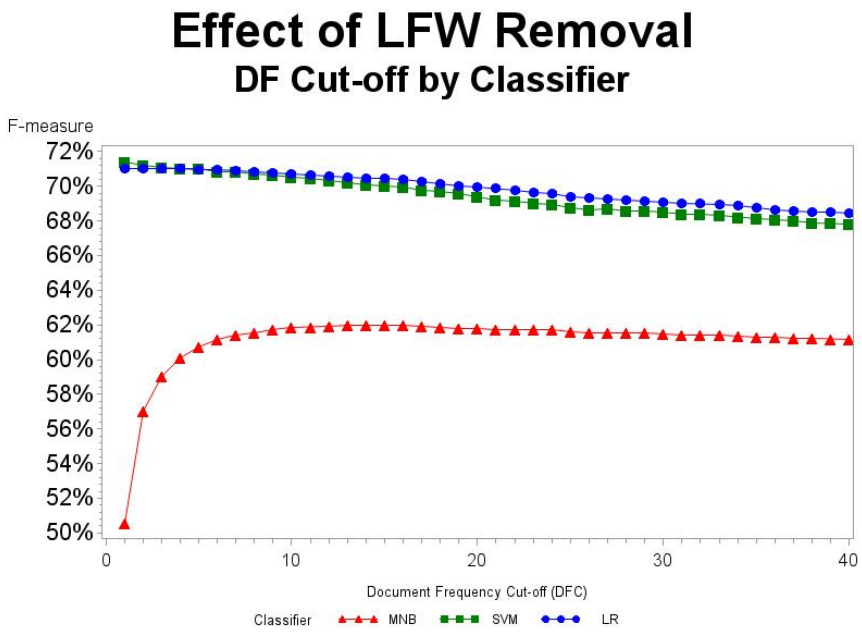

Figure 4.5: Effect of LFW Removal: Overall Classification Performance from DFC Level 1 to 40

Considering that the QISU dataset is a huge corpus with a half-million documents, this study focused on the words that had a DF lower than 10 (i.e., DF1 to DF9 words) and defined as LFWs for the following experiments. The DFC range of interest was set to DFC1 (keeping all words) to DFC10 (removing DF1 to DF9 words). Figures 4.6-4.8 show the overall classification performance as LFWs were continuously being removed by increasing DFC, with one figure for each train-test ratio scenario. Figures 4.6-4.8 demonstrate the effect of removing LFWs: insignificant declining performance for SVM and LR while increasing performance for MNB, consistent with Figure 4.5. The ANOVA test was performed to determine if removing LFWs statistically significantly impacts the classification performance ("DFC"). The results are listed in Table B.1 in Appendix B, showing the significance of the main factor DFC and three-way interaction of Classifier*DFC*Train-Test_Ratio. Thus, separate ANOVA tests were conducted to determine the significance of DFC for each classifier. As Tables B.2-B.4 suggest, removing LFWs did

not have any statistically significant impact on SVM and LR but it did have a significant impact on MNB. Removing the words that occurred more than once significantly increased the classification performance of MNB (see Table B.5 for the result of the post hoc tests). Refer to Table B.6 for the numerical values of the overall classification performance (macro-averaged F-measure) for three classifiers in three train-test ratio scenarios at the DFC level from 1 to 10. Refer to Table B.7 for the numerical values of the overall effect of removing DF1 to DF9 words on the classification performance of three classifiers in three train-test ratio scenarios.

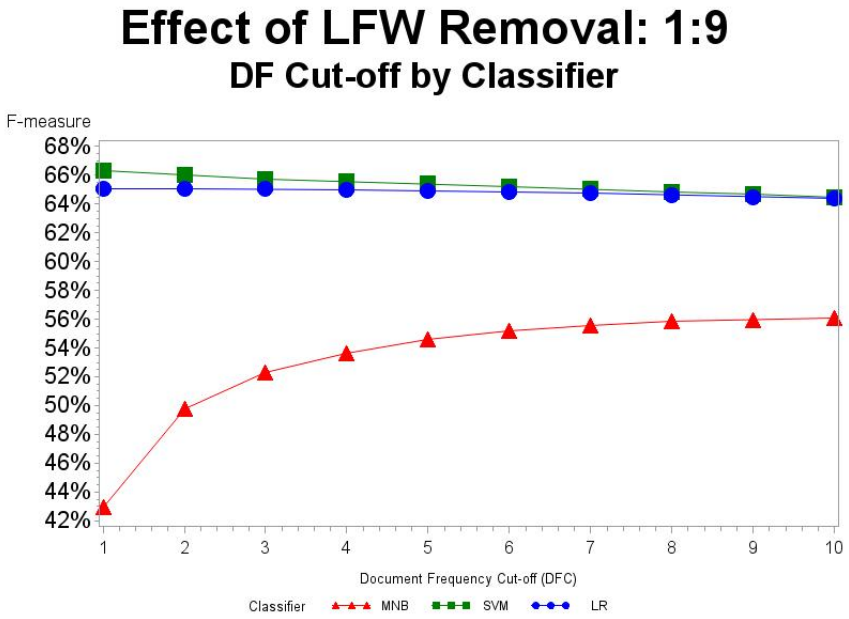

Figure 4.6: Effect of LFW Removal: Overall Classification Performance from DFC Level 1 to 10 at Train-Test Ratio of 1:9

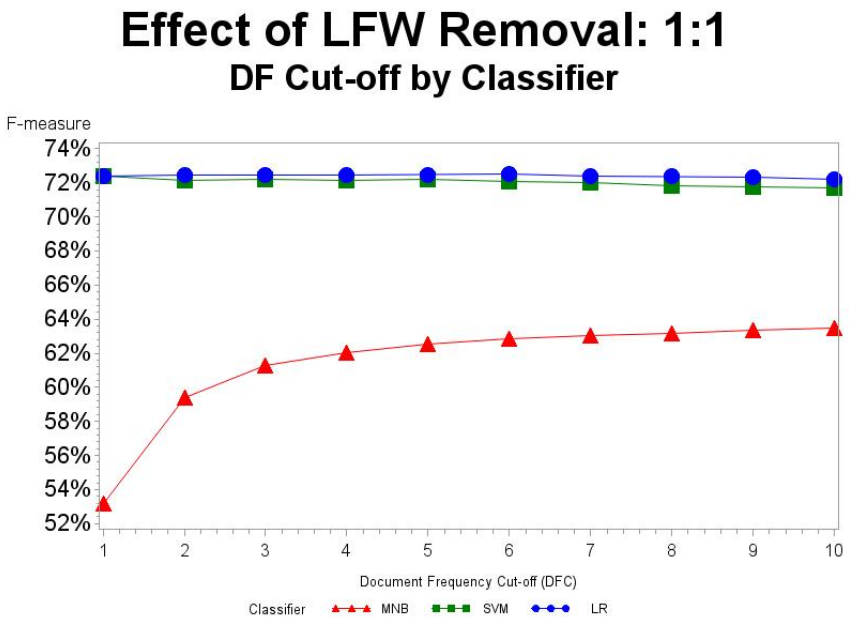

Figure 4.7: Effect of LFW Removal: Overall Classification Performance from DFC Level 1 to 10 at Train-Test Ratio of 1:1

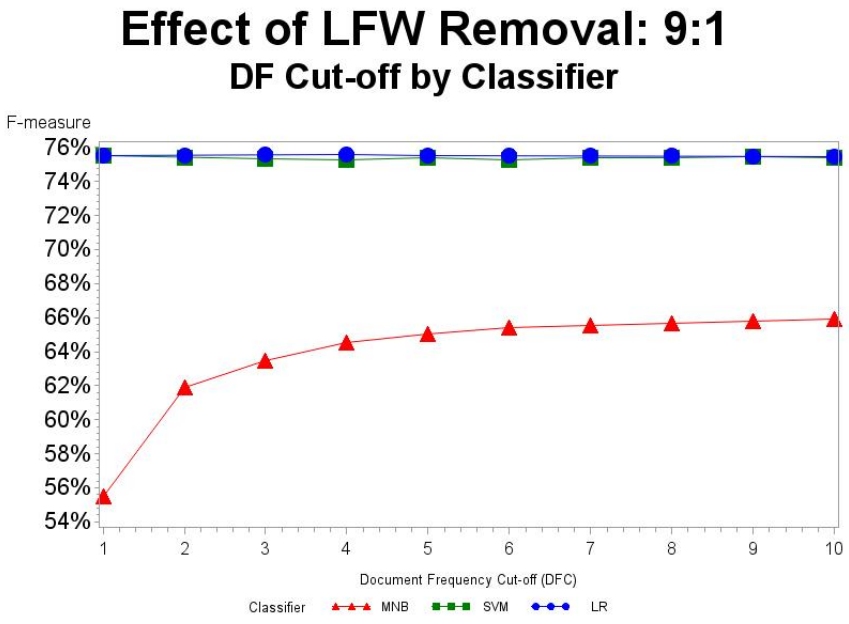

Figure 4.8: Effect of LFW Removal: Overall Classification Performance from DFC Level 1 to 10 at Train-Test Ratio of 9:1

The next step is to explore the effect of LFW removal at different levels of data scarcity, that is, different category sizes and train-test ratios. Three category sizes are small (S), medium (M), and large (L). Considering two extremes and one middle level, three train-test ratios are 1:9 (training data are as little as one-tenth of test data), 1:1 (training and test data are comparable in quantity), and 9:1 (training data are as much as nine times of test data). Figures 4.9-4.11 graphically show the effect of removing LFW on classification performance for three category sizes and three train-test ratio scenarios. For the numerical values of same results, refer to Table B.8 for the overall classification performance for different category sizes at DFC levels from 1 to 10 and Table B.9 for the impact of removing LFWs on the performance for different category sizes at DFC Level 2 (dropping DF1) to 10 (removing DF1 to DF9 words).

Three classifiers showed different effects of LFW removal on the classification performance of different category sizes at different train-test ratio scenarios. For MNB in Figure 4.9, removing LFWs had a significant positive impact on classification performance of small categories, moderate positive impact on medium categories, and insignificant positive on large categories. Removing LFWs had a higher positive effect on small categories when the train-test ratio is larger while medium categories showed an opposite trend.

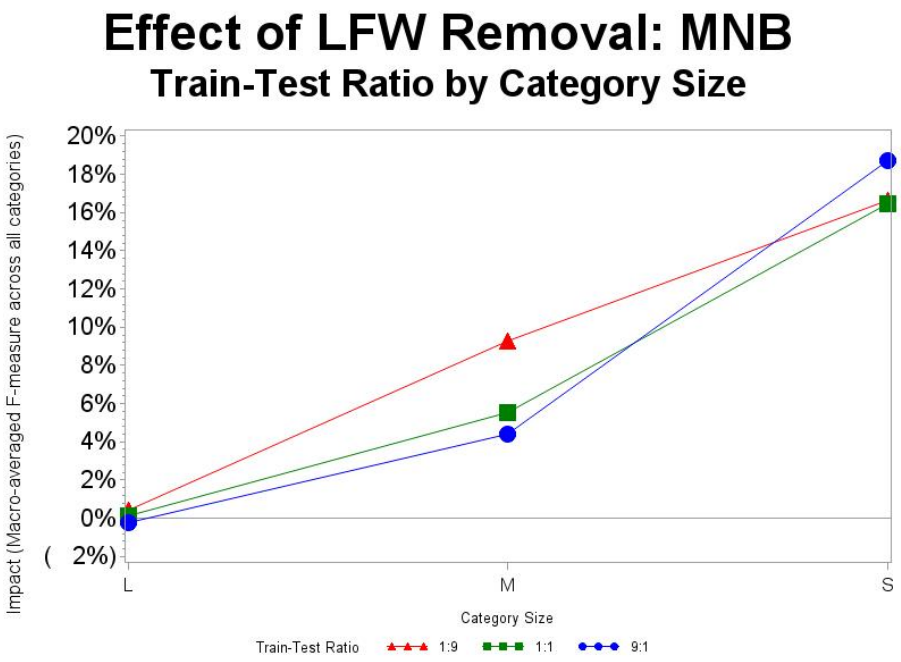

Figure 4.9: Effect of LFW Removal: Category Size by Train-Test Ratio for MNB

For SVM in Figure 4.10, removing LFWs was found to have a moderate negative impact on classification performance of small categories and minor negative impact on medium categories, but slight positive impact on large categories. As the train-test ratio decreased, the effect of removing LFWs became more significant, either increasing positive effect on the classification performance of large categories (0.15% at 9:1; 0.21% at 1:1; 0.32% at 1:9) or increasing negative effect on small categories (-0.32% at 9:1; -1% at 1:1; -2.16% at 1:9).

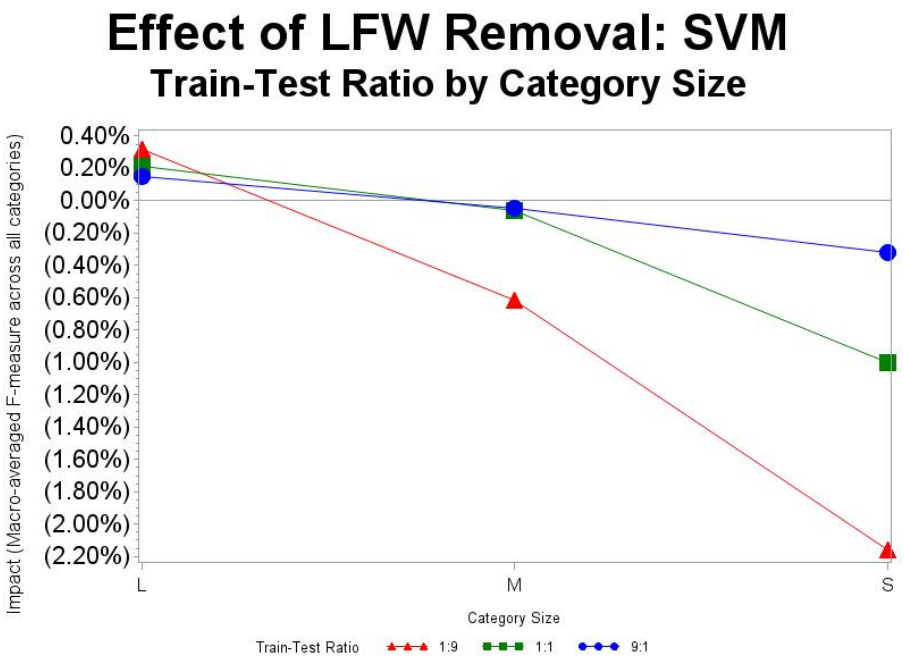

Figure 4.10: Effect of LFW Removal: Category Size by Train-Test Ratio for SVM

For LR in Figure 4.11, removing LFWs had slightly negative impact on the classification performance of large categories and minor negative on medium categories in three train-test data scenarios. However, removing LFWs had the highest negative impact on small categories at the train-test ratio of 1:9 while highest positive impact at the rato of 9:1.

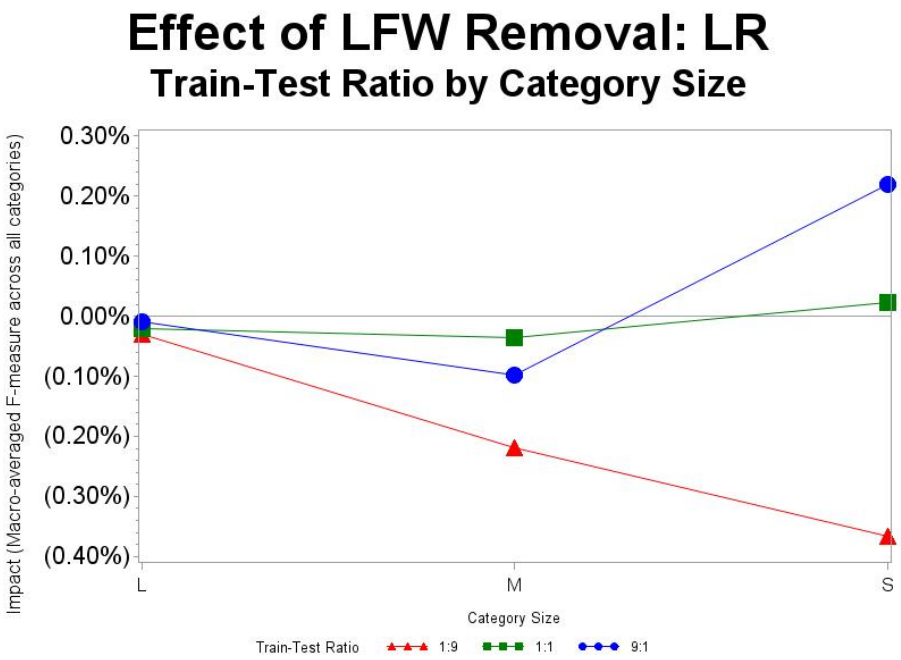

Figure 4.11: Effect of LFW Removal: Category Size by Train-Test Ratio for LR

The effect of removing LFWs was further examined at the category level. Figure 4.12 shows the effect of removing LFWs on the classification performance of each category, from smallest to largest at the *X*-axis. It can be observed that removing LFWs tended to improve the classification of small categories (located in the left chart block that covers FIREARM to FIREFLAME) more when the train-test ratio is at 9:1 than 1:9. The impact differences between two extreme ratio scenarios were significant for the two smallest categories, FIREARM (2%) and PA_PLANT (1.2%). The great improvements in these two smallest categories explained the overall positive effect of LFW removal on small categories at the ratio of 9:1, observed in Figure 4.11.

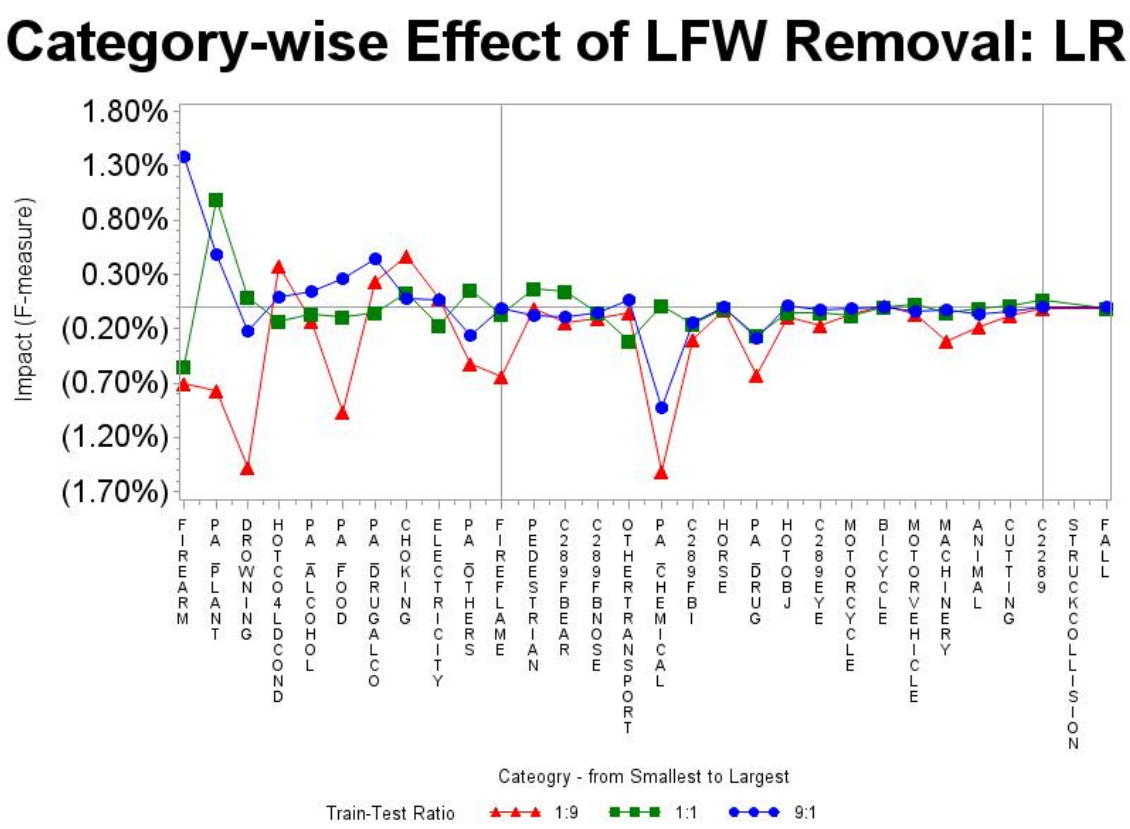

Figure 4.12: Category-wise Effect of LFW Removal on Classification Performance of LR

Transformed Frequency

It was believed that words could be more fairly compared by considering their sampling sizes. In an imbalanced dataset, words that have the same occurrence frequency may not have the same probability. Words from small categories have generally a lower frequency because of the smaller sampling size, and vice versa. Low-frequency words (LFWs) from small categories are not the same as the LFWs from large categories as their sampling sizes are different. In this sense, using the frequency normalized (divided) by category size seems to be a more fair comparison.

The so-called "Transformed Frequency" (TF) of a word w is defined as the sum of the occurrence count normalized by the size of category within which the word w occurs, over all the categories that contain the word w, that is:

$$TransFreq(w) = \sum_{C_i \in C} \frac{count(w; C_i)}{p(C_i)} \tag{4.1}$$

where $count(w; C_i)$ is the number of documents coded as C_i that contain the word w, $P(C_i)$ is the probability of category C_i, and C is a complete set of categories.

One of the hypotheses to test is that LFWs from small categories are more important than the ones from large categories in text classification. A word is said to be important in text classification if its absence negatively impacts the classification performance. The rationale is that large categories already have more information (i.e., more training samples and features) to be trained on than small categories. This study focused on the words that occurred only in one training document (DF1 LFWs). This hypothesis can be tested by examining the change to classification performance due to removing DF1 LFWs from one category at a time. Transformed Frequency (TF) in Equation 4.1 served as a good way to prioritize DF1 LFWs. Words from small categories receive a higher weight than ones from large categories, so LFWs from smaller categories have a higher TF than the ones with the same frequency but from larger categories. Thus, I tested the hypothesis that LFWs from small categories are more important than ones from large categories in text classification by using the TF as a criteria to remove LFWs and examining the resulting performance change. More specifically, the overall effect of removing LFWs on classification performance for different category sizes (S, M, L) were evaluated by examining the performance change corresponding to the category-wise gradual removal of DF1 LFWs by their Transformed Frequency (TF). The rationale is to remove DF1s starting from the largest category, and then continuously removing more DF1 words from the second largest category, and so on, to the smallest category, until all DF1 LFWs are removed.

According to Equation 4.1, the TF of DF1 LFWs is the inverse of the category probability. At the TF Cutoff Level 1, the TF Cutoff is set to the smallest TF of the entire corpus: 3.15369 (i.e., the TF of the DF1 LFWs from the largest category, FALL). Only words that have a TF higher than the cutoff are retained, and thus those DF1 LFWs from FALL category are removed at the TF Cutoff Level 1. At the TF Cutoff Level 2, the DF1 LFWs from the second largest category (i.e., STRUCKBYCOLLISION) along with the ones that were previously removed are then removed because their TF are not higher than the cutoff (3.90267), and so on. Simply put, the DF1 LFWs are removed from the largest category first at TF Cutoff level 1, and then the ones from the second largest category are continuously removed, and so on, until all DF1 LFWs are removed at the TF Cutoff Level 30. Table 4.3 lists the TF of DF1 LFWs for each category and the corresponding TF Cutoff Level that uses the value of TF as a cutoff, from largest to smallest category.

Table 4.3: Transformed Frequency (TF) of DF1 LFWs and TF Cutoff Level

Category (Category Size)	% of Cases	Transformed Freq. (TF) of DF1 words from this category as Cutoff	TF Cutoff Level
FALL (L)	31.71%	3.15369	1
STRUCKCOLLISION (L)	25.62%	3.90267	2
C289 (M)	9.23%	10.83512	3
CUTTING (M)	7.10%	14.08780	4
ANIMAL (M)	4.01%	24.96483	5
MACHINERY (M)	3.50%	28.56089	6
MOTORVEHICLE (M)	2.99%	33.47397	7
BICYCLE (M)	2.63%	38.06588	8
MOTORCYCLE (M)	2.37%	42.23827	9
C289EYE (M)	2.02%	49.51496	10
HOTOBJ (M)	1.94%	51.64555	11
PA_DRUG (M)	1.12%	89.60601	12
HORSE (M)	0.76%	132.14651	13
C289FBI (M)	0.75%	132.65281	14
PA_CHEMICAL (M)	0.58%	173.64622	15
OTHERTRANSPORT (M)	0.56%	179.53371	16
C289FBNOSE (M)	0.51%	195.94732	17
C289FBEAR (M)	0.42%	238.39566	18
PEDESTRIAN (M)	0.38%	260.46933	19
FIREFLAME (S)	0.32%	315.63184	20
PA_OTHERS (S)	0.30%	333.40074	21
ELECTRICITY (S)	0.26%	381.43305	22
CHOKING (S)	0.23%	432.36407	23
PA_DRUGALCO (S)	0.19%	527.03864	24
PA_FOOD (S)	0.17%	574.82886	25
PA_ALCOHOL (S)	0.11%	926.11317	26
HOTCOLDCOND (S)	0.09%	1087.17633	27
DROWNING (S)	0.09%	1111.33580	28
PA_PLANT (S)	0.04%	2273.18687	29
FIREARM (S)	0.01%	7032.67188	30

The impact of removing DF1 LFWs by TF was evaluated by the difference in the macro-averaged F-measure between keeping all words and removing words at a given TF Cutoff level. Figure 4.13 shows the impact on the macro-averaged F-measure of small, medium, and large categories for three classifiers (LR, SVM, MNB) in three train-test ratio scenarios (1:9, 1:1, 9:1) as the cutoff level of Transformed Frequency (TF) increases.

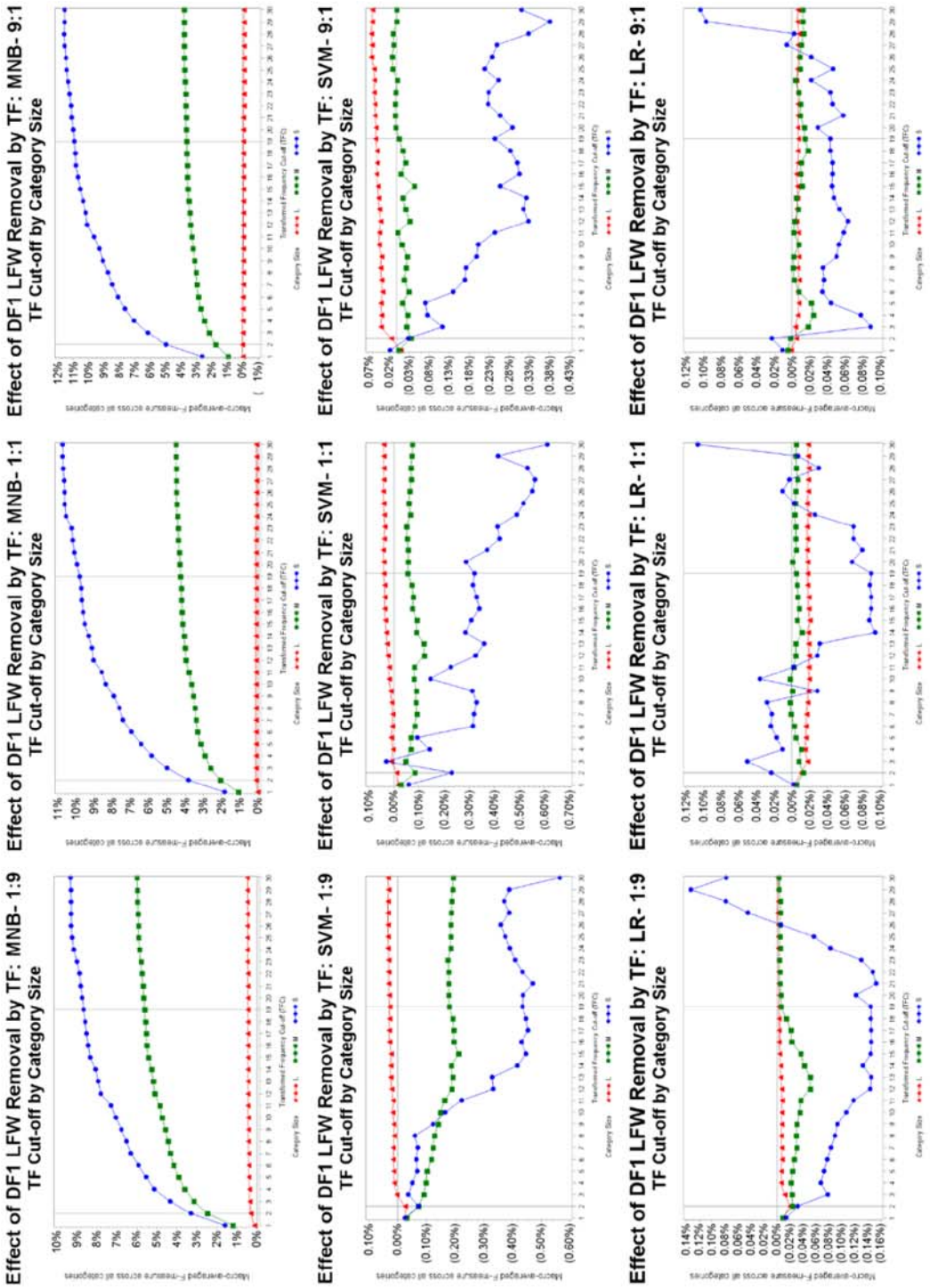

Figure 4.13: Impact of Removing DF1 LFWs by Transformed Frequency for MNB, SVM, LR in Three Train-Test Ratios

Due to their distinct patterns, the effects of removing DF1 LFWs by TF are discussed for each classifier in the following:

LR: For all three train-test ratio scenarios, removing DF1 LFWs by TF, for LR, was found to have no or slightly negative impact on F-measure of medium and large categories (within the range of +/- 0.05%), but has a trend of impact first going negative then turning positive for small categories. During the initial phase of declining performance, the average negative impact is -0.15% for the train-test ratio 1:9, -0.1% for 1:1, and -0.07% for 9:1. The smaller train-test ratios had more severe negative impact. After starting to remove DF1 LFWs from small categories (TF Cutoff Level 20), the effect of LFW removal started to turn positive and had a general increasing until all DF1 LFWs were removed at TF Cutoff Level 30 (the positive impact reached + 0.1%).

SVM: Overall, dropping DF1 LFWs, for SVM, had a slight positive impact on the performance of large categories (+0.05%), minor negative impact on medium categories (up to -0.2%), and increasingly negative impact, although with fluctuation, on small categories (up to -0.6%). The negative impact on small categories was found more severe when training examples are limited than when sufficient in relation to the test examples (up to -0.6% at the ratio of 1:9 and -0.4% at 9:1).

MNB: Removing DF1 LFWs by TF, for MNB, had no or minor positive impact on the classification performance of large categories (up to +0.5%), moderate positive impact on medium categories (up to +6%), and significant positive impact on small categories (up to +12%). Dropping LFWs improved MNB to classify small categories up to 12% at the train-test ratio of 9:1, followed by 11% at 1:1, and 9.5% at 1:9. Dropping DF1 LFWs achieved the most improvement when training examples were sufficient (at 9:1) because LFWs were more likely to be noise.

Next I intend to provide an explanation for the U-shaped trend of impact that was found in the small categories for LR when removing LFWs by TF (i.e., going downward then upward). Figures 4.14-4.16 show the impact on the macro-averaged F-measure for each small category as the TF Cutoff Level increases, i.e.,

continuously removing DF1 LFWs from largest (TF Cutoff Level 1) to smallest categories (TF Cutoff Level 30). The absolute values are not of interest here, but rather the overall trend that can potentially explain the U-shaped impact on the small categories for LR.

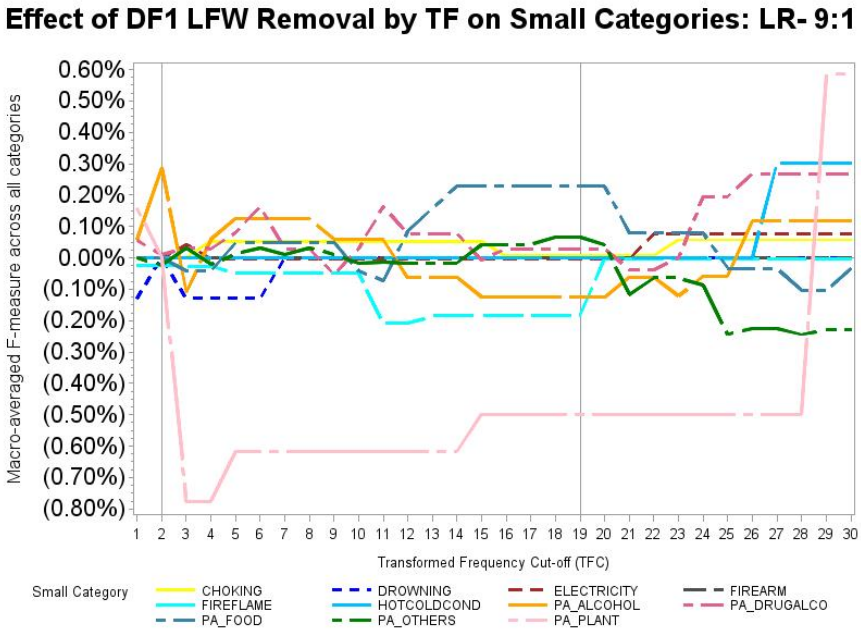

Figure 4.14: Impact of Removing DF1 LFWs by Transformed Frequency on Small Categories for LR at Train-Test Ratio of 9:1

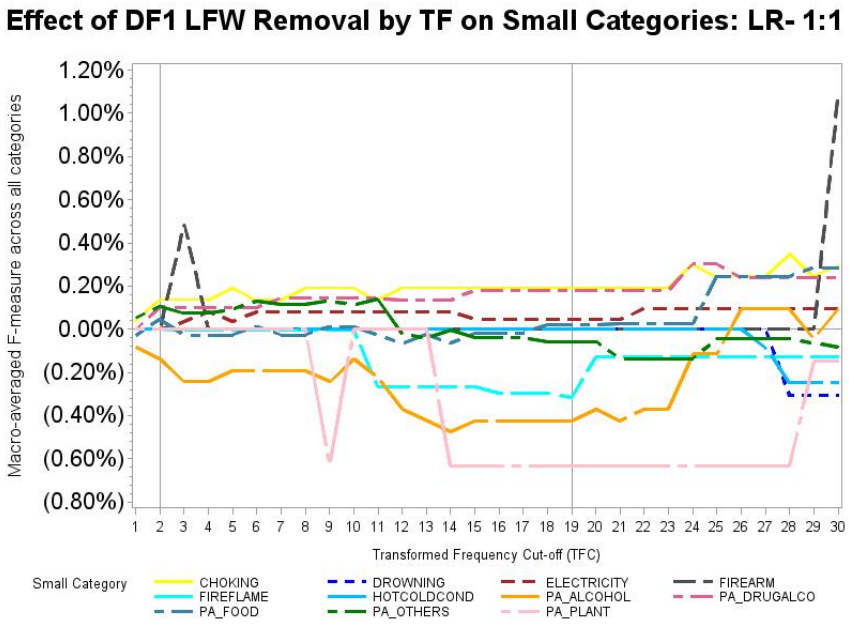

Figure 4.15: Impact of Removing DF1 LFWs by Transformed Frequency on Small Categories for LR at Train-Test Ratio of 1:1

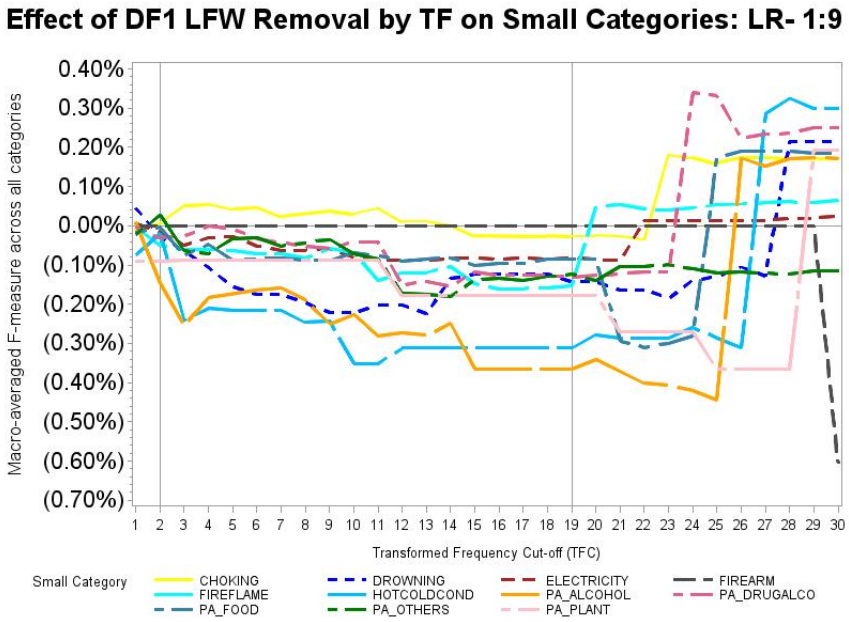

Figure 4.16: Impact of Removing DF1 LFWs by Transformed Frequency on Small Categories for LR at Train-Test Ratio of 1:9

Here, the focus in Figures 4.14-4.16 is the change of classification performance at the TF Cutoff Level right after the DF1 LFWs from the same category are removed and how the removal of same-category DF1 LFWs would affect the classification performance of each small category. From Figures 4.14-4.16, the same-category-DF1-LFWs-removal effect can be classified into three types:

1. Performance decreases sharply after same-category-DF1-LFWs are removed and stays at similar level afterwards (Effect-D)

2. Performance increases sharply after same-category-DF1-LFWs are removed and stays at similar level afterwards (Effect-I)

3. Performance fluctuates in no or limited association with the removal of DF1 LFWs (Effect-N)

Table 4.4 lists the corresponding effects for each category, in an order of smallest to largest category, in three train-test ratio scenarios.

Table 4.4: Distribution of Same-Category-DF1-LFWs-Removal Effects for Categories

Category (Category Size)	**% of Cases**	**1:9**	**1:1**	**9:1**
FIREARM (S)	0.01%	D	I	N
PA_PLANT (S)	0.04%	I	I	I
DROWNING (S)	0.09%	I	D	N
HOTCOLDCOND (S)	0.09%	I	D	I
PA_ALCOHOL (S)	0.11%	I	I	I
PA_FOOD (S)	0.17%	I	I	D
PA_DRUGALCO (S)	0.19%	I	I	I
CHOKING (S)	0.23%	I	N	I
ELECTRICITY (S)	0.26%	I	I	I
PA_OTHERS (S)	0.30%	N	D	D
FIREFLAME (S)	0.32%	I	I	I
PEDESTRIAN (M)	0.38%	I	N	N
C289FBEAR (M)	0.42%	I	N	D
C289FBNOSE (M)	0.51%	N	N	N
OTHERTRANSPORT (M)	0.56%	I	N	N
PA_CHEMICAL (M)	0.58%	I	N	D
C289FBI (M)	0.75%	I	D	N
HORSE (M)	0.76%	N	N	N
PA_DRUG (M)	1.12%	D	N	D
HOTOBJ (M)	1.94%	N	N	N
C289EYE (M)	2.02%	D	N	N
MOTORCYCLE (M)	2.37%	N	N	N
BICYCLE (M)	2.63%	N	N	N
MOTORVEHICLE (M)	2.99%	D	N	N
MACHINERY (M)	3.50%	N	I	N
ANIMAL (M)	4.01%	D	N	N
CUTTING (M)	7.10%	N	N	N
C289 (M)	9.23%	I	I	N
STRUCKCOLLISION (L)	25.62%	N	N	N
FALL (L)	31.71%	N	N	N

Removing LFWs may result in the loss of discriminatory features for properly classifying some categories, and thus a performance drop can be expected (Effect-D). Table 4.4 suggests that the Effect-D can happen to any size of category and there is no obvious occurrence pattern. In addition, the Effect-N tended to occur in larger categories because they are often equipped with a great amount of features and thus, less likely to be influenced by the removal of LFWs.

As for the Effect-I, Table 4.4 indicates that smaller categories were more likely to experience the Effect-I, which describes the situation that the classification performance is improved right after removing the DF1 LFWs from the same category. Note that the category C289, as a large category, was also observed to have the Effect-I. C289 is a category for cases with other and unspecified external causes, and thus is a comprehensive category that combines many small, diverse categories. This explains why C289 is prone to have Effect-I, like other small categories.

The finding that, for LR, the classification performance of smaller categories was likely to increase when same-category LFWs were removed (Effect-I) corresponds to the U-shape performance trend for small categories, where the classification performance of small categories started to improve when DF1 LFWs from small categories were starting to be removed.

The Effect-I and dramatic increase of classification performance when same-category LFWs were removed, observed in the small categories for LR, could be an artifact of overfitting. The literature has suggested that LR is prone to overfitting when training examples are limited in quantity. An overfitted model often gives a high weight on extremely low frequency features and relies heavily on those rare features that may not occur again, and thus, fails to be trained to predict properly. The performance increase due to the removal of same-category DF1 LFWs might be because the removal of those extremely rare features (DF1 LFWs) alleviates the overfitting problem of the model by redistributing the weight more evenly to other higher frequency, more representative features.

Different availability of training cases in relation to test cases (train-test ratio) also affects the distribution of different effects. Table 4.5 lists the category counts for three types of same-category-DF1-LFWs-removal effects in three train-test ratio scenarios (1:9, 1:1, 9:1).

Table 4.5: Distribution of Same-Category-DF1-LFWs-Removal Effects for Three Train-Test Ratio Scenarios

Removal Effect	1:9	1:1	9:1
Effect D(ecrease)	5	4	5
Effect I(ncrease)	15	9	7
Effect N(o association)	10	17	18

As Table 4.5 shows, the distribution of Effect-D had no difference among three train-test ratio scenarios. This finding is consistent with the observation that the Effect-D had no occurrence tendency for any size of category. When training cases were limited (ratio of 1:9), the Effect-I was more likely to occur while the Effect-N was not. On the other hand, when training cases became more available (ratio of 9:1), the Effect-I started to diminish whereas the Effect-N became more frequent. The results, again, correspond with the observation that the Effect-I was more likely to occur in small categories while the Effect-N was more likely to occur in large categories.

To sum up the effect of removing same-category LFWs, when training cases were sufficient (large category or train-test ratio), removing LFWs from the same category was less likely to influence the classification performance (Effect-N). Also, the positive effect of same-category-LFWs-removal (Effect-I) was more likely to occur when training cases were limited (small category or train-test ratio). This result could be explained as an artifact of overfitting.

Given the discussion above, Null Hypotheses 2 to 4, associated with LFWs, can be proved or disproved accordingly:

Null Hypothesis 2: *Low-frequency words are NOT important for text classification of injury data because the absence of them does not influence the classification performance.*
Alternative Hypothesis 2: *Low-frequency words are important for text classification of injury data because the absence of them can deteriorate the classification performance.*

Null Hypothesis 2 was not rejected and thus Alternative Hypothesis 2 was not accepted for MNB as removing LFWs improved the classification performance of MNB. As a result, Null Hypothesis 3 and 4 in the following are not tested for MNB.

Null Hypothesis 2 was rejected and Alternative Hypothesis 2 was accepted for SVM as it was observed that removing LFWs negatively impacted the classification performance of SVM.

Null Hypothesis 2 was rejected and Alternative Hypothesis 2 was accepted for LR as removing extremely LFWs (DF <3) improved the classification performance but removing other LFWs deteriorated the performance.

Null Hypothesis 3 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are important regardless of category sizes in text classification of injury data because the absence of them causes similar negative impact on the classification performance for any category sizes.*
Alternative Hypothesis 3 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are more important for small categories than large categories in text classification of injury data because the absence of them can cause more negative impact on the classification performance of small categories than large categories.*

Null Hypothesis 3 was falsified and Alternative Hypothesis 3 was accepted for SVM because removing LFWs negatively impacted the classification performance of small categories while slightly improved the performance of large categories.

Null Hypothesis 3 was not falsified and Alternative Hypothesis 3 was not accepted for LR as removing LFWs that occurred only once or twice (DF <3 words) slightly improved the overall classification performance of small categories (+0.08%). By exploring the effect of removing DF1 words from the same category, it was observed that the performance of many small categories increased sharply after same-category DF1 words are removed. As LR is known to be prone to overfitting with scarce training data, this positive impact of removing LFWs can be explained as the overfitting problem being alleviated. The removal of LFWs helped the LR to assign weights more evenly and properly and thus the model can rely on more frequent and discriminatory features, rather than the rare features that have very low chance to occur again in a test set.

Null Hypothesis 4 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are important for text classification regardless of sizes of training and test datasets, because the absence of them causes similar negative impact on the classification performance in any ratio of training data to test dataset sizes.*
Alternative Hypothesis 4 (Given that Null Hypothesis 2 is rejected): *Low-frequency words are important for text classification when the training dataset size is limited rather than sufficient relative to the test dataset size, because the absence of them can cause more negative impact on the classification performance when the training dataset size is far smaller than the test dataset size, compared to when the training dataset size is far larger than the test dataset size.*

Null Hypothesis 4 was falsified and Alternative Hypothesis 4 was accepted for

SVM and LR. To evaluate the effect of removing LFWs on classification performance at different availability levels of training and test data, the ratio of size of training data to test data, called "train-test ratio," was considered. Three train-test ratio scenarios were two extremes and one in middle: 1:9, 1:1, and 9:1. Removing LFWs was found to have a negative impact on the classification performance for SVM and LR, though positive for MNB. At the train-test ratio of 1:9 (limited training data to predict huge test data), the overall impact, either positive or negative, was more significant than 9:1. For MNB, removing LFWs had more positive impact at the ratio of 1:9 (11.4%) than ratio of 9:1 (9.33%). For SVM and LR, on the other hand, removing LFWs had more severe negative impact at the ratio of 1:9 (-1.12% for SVM and -0.27% for LR) than ratio of 9:1 (-0.13% for SVM and +0.02% for LR). The findings with/about/regarding SVM and LR falsified Null Hypothesis 4 and Alternative Hypothesis 4 was then accepted.

Overfitting tends to occur when training examples are limited and the model tries to fit every feature, especially the extremely rare ones. Thus, extremely low frequency features are given a higher weight than they should have, which makes the model biased in favor of rare features. An over-fitted model often relies heavily on those infrequent features whose recurrence in a test set is questionable. Thus, it often predicts well on the training set but poorly on the test set. One way to avoid the overfitting problem is to create a smaller but denser feature space with more representative and discriminatory features to train a machine learning model. Many researchers have developed a variety of word normalization methods and feature selection methods to reduce the dimensionality of feature space for a better machine learning environment by grouping or prioritizing features. In the following chapters, I will discuss these methods and their limitations in more detail and propose the word grouping method that addresses these limitations for potentially better classification performance.

5. INTRODUCTION TO TYPE M+S GROUPING METHOD

5.1 Background and Limitations of Stemming and Lemmatization

For grammatical reasons, words have different forms, inflected and derivational forms, which are products of two types of word formation processes. Inflection is a process that grammatical variants of the same words are formed. Inflected forms are used to signify a change in number (singular or plural) or tense (past, present, future). For example, most English plural nouns are inflected with an inflectional affix "-s" (as in "car" and "cars"). English verbs are often inflected with an inflectional affix "-s" for the third person singular present tense, "-ed" for the past tense, and "-ing" for the present participle (as in "talks", "talked", and "talking").

Derivation, on the other hand, is a process of creating a new word, a distinct lexeme, on the basis of an existing word. A lexeme is a unit of meaning. In English, for example, *run*, *runs*, *ran*, and *running* are forms of the same lexeme. English derivational affixes often change either the part of speech or meaning of the affected word. Take the adjective "happy" as an example. The addition of a derivational suffix "-ness" produces the word "happiness", which changes the part of speech from adjective to noun. When applying a derivational prefix "un-" to the root word happy, a new word "unhappy" is created with an opposite meaning. The derivational prefix "un-" here serves as a "negative prefix" and gives a negative connotation to words. Despite the changed meaning, the derived word still has the meaning, although being negated, related to its base form.

Derivation and inflection can be contrasted with each other in the sense that inflection produces grammatical variants of the same word without changing its meaning, whereas derivation can produce a new word with a different, but related,

meaning. However, as long as words are transformed from their base form, their essential meaning is still related to their base form.

In information retrieval and natural language process, a word normalization method is often applied as a text preprocessing task. The purpose is to reduce inflectional forms and sometimes derivationally related forms of a word to its common root or base form, which in turn, leads to a smaller, denser feature space model and improves the statistical robustness of discriminatory concepts.

Two common word normalization or grouping methods are stemming and lemmatization. This study used the Porter Stemmer (Porter, 1980) and WordNet Lemmatizer, supported by the Python NLTK package. Their demos are available online at http://text-processing.com/demo/stem/ (Bird et al., 2009).

The Porter Stemmer is one of the most commonly used stemmer in practice. Stemmers rely on a collection of heuristic rules to remove the ending of words in the hope of transforming words into their stem (Manning et. al, 2008). For example, Porters algorithm (Porter, 1980) consisted of 5 sequential phases of word reductions. Each phase has various conventions for rule selection. Take the first phase as an example. The convention is used within the following rule groups:

Table 5.1: Two-Way Contingency Table of Term and Category

Rule	**Example (word → stem)**
SSES → SS	caresses → caress
IES → I	Ponies → poni
SS → SS	caress → caress
S→	cats → cat

Note that a stem may not be a real word. The exact stemmed form does not matter because, more importantly, words with the same stem can be grouped and

merged. Stemming is rather aggressive in grouping words, which often results in an improved recall at the cost of precision. An example from Manning et. al. (2008) is that the Porter's algorithm stems all of the following words: *operate operating operates operation operative operatives operational* to *oper* .

One disadvantage of stemming is that words may not be mapped properly to the stem if they are misspelled. For example, the words electricity and electrical are stemmed to *electr* while their misspelled version *electrocity* and *electrial* are stemmed to *electroc* and *electri*, rather than the ideal one *electr*.

Lemmatization is another word normalization method that utilizes a dictionary and full morphological analysis to identify the lemma (i.e., base or dictionary form) for each word. The WordNet Lemmatizer makes use of the WordNet lexical database to look up lemmas. Since a word may have multiple word senses and parts of speech (POS), the WordNet Lemmatizer takes POS into account; however, if the POS is not specified, it is a noun by default in the program. The use of WordNet dictionary improves the precision, but also decreases the recall of mapping words and constrains itself to the following two limitations:

1. Lemmatizer is incapable of handling misspellings or domain-specific terms. The WordNet Lemmatizer looks up the word in the WordNet database and only transforms the word into its lemma when it is in the database. Thus, the lemmatization becomes less effective when the operational corpus is free-text in nature (prone to misspellings) or related to a specific domain (such as a drug name), which are indeed quite common in real-world practice.

2. Lemmatizer requires a correct POS as an argument to properly lemmatize a word. However, annotating every word with a proper POS tag in a text corpus for lemmatization is often impractical due to the large vocabulary involved. Dealing with thousands or tens of thousands of words is common in practice. Words often have multiple parts of speech at different times and can mean differently in different contexts. Differentiation between the

> POS tags for ambiguous words is still challenging. The automated POS tagging algorithm, although available, is not reliable when the training and operational data have differences in topic, epoch, or writing style. The development of a POS tagger involves a POS-annotated training corpus that often requires tremendous manual effort. For example, the Stanford Part-of-Speech Tagger (online demo: http://nlp.stanford.edu:8080/corenlp/process) uses the annotated corpora of Penn Treebank as their tagset. The Stanford team claimed that their POS Tagger has 97% per-token accuracies (punctuation marks count) and 56% sentence accuracies (Toutanova, Klein, Manning, & Singer, 2003). However, accuracies can drop significantly when the operational test set does not have the similar topic, domain, or writing style as the annotated training set (Manning, 2011). Thus, the limited availability of correct POS information for an operational dataset would negatively influence the effectiveness of the lemmatizer.

Despite the limitations above, stemming and lemmatization are still commonly applied as text preprocessing tasks for grouping low-frequency words (LFWs) and normalizing vocabulary. The traditional methods normalize or merge words with the same stem or base form assuming that they carry the similar semantic meaning and predict the same category, however, without examining their actual predictive categories.

An ideal word normalization method should group words that have the same hypernym and predict the same category into a united form as the representative. A hypernym is a word that names a broad category that includes other words. For example, “musical instrument” is a hypernym of “guitar” and “piano”. The goal is to group LFWs, reduce the vocabulary size, improve the statistical robustness of discriminatory concepts, and thus improve the performance of statistical text analysis.

However, many words that have a similar semantic meaning or hypernym do not come from the same base form. Those same-hypernym words can be spelled

differently but mean similarly. Traditional normalization methods are designed to group same-hypernym words with similar spellings but they are incapable of grouping them if they have different spellings.

Furthermore, the traditional classification only considers the words occurred in the training set while ignoring the words that only occurred in the test set. However, those words that occur only in the test set but not in the training set can be valuable if they carry the meaning that is essential to classify a category. In the sense that they have zero occurrence in the training set, the words only seen in the test set can be viewed as a type of LFW.

5.2 Proposed Word Grouping Method

In this study, I proposed the so-called Type M+S Grouping method that aimed to address the limitations of stemming and lemmatization. More specifically, the proposed grouping method was expected to:

1. have better capability of handling misspellings, domain-specific terms, and same-hypernym words with a different spelling,

2. be more grounded and less aggressive because the words that predict different categories will not be merged, and

3. be capable of identifying and utilizing discriminatory words that only occurred in the test set, which are ignored in traditional text analysis.

Type M+S Grouping method, as its name suggest, is comprised of two types of grouping strategies based on the nature of the words each method identifies and groups. Two types of important LFWs are identified for classification. The Type-M (Morphological) LFWs are the misspellings or morphological variants of strong predictor words whereas the Type-S (Semantic) LFWs are the rare terms whose hypernym or high-level concept is essential to classify a certain category.

The rationale behind the Type M+S Grouping method is simple. To address the incapability of stemming and lemmatization to handle misspellings and domain-specific words, the proposed M+S Grouping method considers all words that occurred in either training or test sets. This way, the potentially discriminatory words that only occurred in the test set are also taken into account. To avoid grouping words that predict a conflicting category, their predictive strength or association with categories is considered. The idea is not to group words with conflicting predictive categories. The essence of word grouping or normalization method is to identify and group words with the same hypernym or similar meaning. Words with a similar spelling or form tend to share a base form and carry a similar or related meaning. Words that are unrelated in form can also be related in meaning. Thus, I believe that, by measuring the morphological and semantic similarity of words along with the categories they tend to associate or predict, words can be properly grouped.

There are two types of grouping strategy: mapping or tagging. Mapping is to replace the original word with another word whereas tagging is to add another word to the narrative that contains the original word. Mapping is more aggressive than tagging because mapping removes the original word and uses a representative word while tagging keeps the original word but adds the extra information. The traditional word normalization methods such as stemming and lemmatization use the mapping strategy by mapping the inflected or derivational words to their stem or base form. In this study, I grouped Type-M LFWs with the mapping strategy as the traditional methods do while also exploring the two options for grouping Type-S LFWs.

5.2.1 Overview of Type-M Morphological Grouping

As a part of Type M+S Grouping Method, the Type-M Mapping Method was proposed to group same-hypernym words that have a similar spelling and non-

conflicting predictive category. The method takes the words in training and test sets into account, and thus can deal with misspelling and domain-specific words which stemming and lemmatization fail to handle. This method also utilizes the coefficient matrix of a trained linear classifier to indicate of the predictive strength of words for categories. The category that has the highest coefficient among all categories is designated as the predictive category for a word. Type-M Mapping groups words with similar spelling and predictive category, and thus is considered to be more conservative and grounded compared to stemming and lemmatization.

The effect of Type-M Mapping was evaluated by the performance difference between mapping and non-grouping. The non-grouping was the standard classification based on the feature space of entire vocabulary without any grouping strategy. The classification performance was measured by the macro-averaged F-measure across all categories.

In Chapter 6, I elaborate the Type-M Mapping Method in more detail, report the effect of Type-M Mapping on classification performance while testing the effectiveness of the coefficient matrix of three linear classifier (MNB, SVM, LR) as an indicator of predictive category, and discuss the results of the ANOVA test and effect of Type-M Mapping.

5.2.2 Overview of Type-S Semantic Grouping

In addition to Type-M Mapping which groups same-hypernym (similar-concept) words with similar spelling, Type-S Grouping aims to group same-hypernym words regardless of spellings.

Similar to stemming and lemmatization, Type-M Mapping assumes that words with similar root or base form have the similar meaning and thus can group words according to their spellings without knowing their meaning. However, grouping same-hypernym words with different spellings require the knowledge of word semantics and meaning. Using a dictionary such as the WordNet Database may

provide the well-defined semantics, but a proper lookup requires the specificity of word sense or part of speech from users. Furthermore, a dictionary like WordNet often fails to handle words that are not indexed (such as misspellings or domain-specific words). The literature has demonstrated that the capability of statistical semantics in capturing the linguistic features of human natural language, both semantically and syntactically. For example, one of the most promising statistical semantics measures called Word2Vec, developed by Google in 2013, was found capable of addressing some of the semantic relations (such as capital city) or syntactic relations (such as word tense). Refer to Section 2.4.4 for more details on Word2Vec.

In Chapter 7, I report the feasibility test of using statistical semantics to identify same-hypernym words with similar meaning for the purpose of improving classification performance. In Section 7.1, I introduce two types of statistical semantics (i.e., correlational and distributional semantics) and the proposed Semantic Data Mining Method that utilizes the statistical semantics to identify words that indicate the similar concept as the pre-selected seedword. The effectiveness of statistical semantics along with the purposed method is discussed based on the level of accuracy of identifying injury agents (drug name) from injury narratives of drug poisoning and allergy. The best-performing statistical semantics is identified and used for the Type-S Semantic Grouping Method in later sections.

In Section 7.2, I discuss the results of an exploratory study for Word2Vec, the statistical measure proved superior in Section 7.1. The purpose of this study is to understand the mechanism of Work2Vec in terms of how statistical similarity are quantified and similar words are ranked. After that, I report the effectiveness of the Semantic Grouping Method, which utilizes the Word2Vec statistical semantics to group same-hypernym words as the text pre-processing task for classification. I also discuss the performance differences between two grouping strategies (tagging and mapping) and the improvement due to different levels of manual reviews involved in Type-S Grouping.

As the proposed Type-S Grouping Method requires seedwords as an argument to identify words that have the same hypernym as these seedwords, domain knowledge is often required for properly selecting seedwords that are indicative of certain categories. In order to eliminate the effort of manually selecting seedwords, I report the effectiveness of using classic feature selection methods to identify discriminatory concepts for Type-S Mapping (proved superior to tagging in Section 7.2) in Section 7.3.

6. TYPE-M MORPHOLOGICAL MAPPING METHOD

As noted in Chapter 5, this chapter is dedicated to the Type-M Mapping Method as the first part of the proposed Type M+S semantic grouping method. The purpose of Type-M Mapping is to group words that share similar meaning and spelling with an ultimate goal of improving the classification performance with a smaller, denser, and more representative feature set.

In the following, I introduce the Type-M Mapping method in Section 6.1, describe the experimental design in Section 6.2, discuss the results in Section 6.3, and summarize the major findings in Section 6.4.

6.1 Type-M Mapping Method

The goal of the proposed Type-M Mapping method is to group similar-concept words that have similar spelling when they do not predict a conflicting category. A word is said to predict a category if its presence is a strong indicator of that category. Such association is often called "discriminatory power" or "predictive strength" in text classification. In order to achieve the goal of Type-M Mapping, the measure of spelling (or morphological) similarity and predictive strength are required and described in the following:

1. Morphological similarity:

   Morphological similarity is the similarity of spelling or form between two words. The measure of the morphological similarity for a word pair can be based on their character n-gram similarity. In the field of computational linguistics, a character n-gram is a contiguous sequence of n letters from a given string of text. In this study, I used the NGram Module in Python

("Python NGram," 2007) to algorithmically quantify the similarity between two words based on the following equation:

$$n-gram\ similarity = \frac{a^e - d^e}{a^e} \tag{6.1}$$

where a is the total number of the distinct n-grams across two words, d is a number of n-grams that are not shared by two words, and e is a tuning parameter (float in 1.0 to 3.0) to increase the similarity of shorter word pairs. Given e=1 by default, the n-gram similarity is the percentage of all possible distinct n-grams that are indeed shared by two words. This study used $n = 2$ and $e = 1$ in the Python's NGram module.

2. Predictive strength:
   Previous research has shown that the coefficient matrix of a linear classifier signifies the importance of words for classifying a category. The word that receives a higher coefficient for a certain category is more important for classifying that category than any others with a lower coefficient.

The procedure of conducting the Type-M Mapping Method is listed in the following and showed in Figure 6.1:

1. Prepare:
   (a) Alphabetical-sorted vocabulary list that include words occurred either in a training or test set
   (b) Document frequency (DF)
   (c) Predictive strength table – coefficient matrix of a trained classifier
2. For each word in the alphabetical-sorted vocabulary list:
   2.1. If the word has an extremely high n-gram similarity score (> 0.8) with the next word in the vocabulary list

2.1.1. Map the word with a lower DF to the word with a higher DF (merge tag word)

2.2. If the word is a strong predictor word (i.e., the coefficient meets a given threshold: -10 for MNB and 0.2 for SVM and LR):

2.2.1. Create the so-called "merge list" that includes the neighboring words in a window of 20 that have at least a moderate n-gram similarity (> 0.5), while excluding the ones that are strong predictors of other categories

2.2.2. Map the words in the merge list to the word with the highest DF (merge tag word)

3. Group words with the same merge tag and create a word-list that organizes words for each merge tag

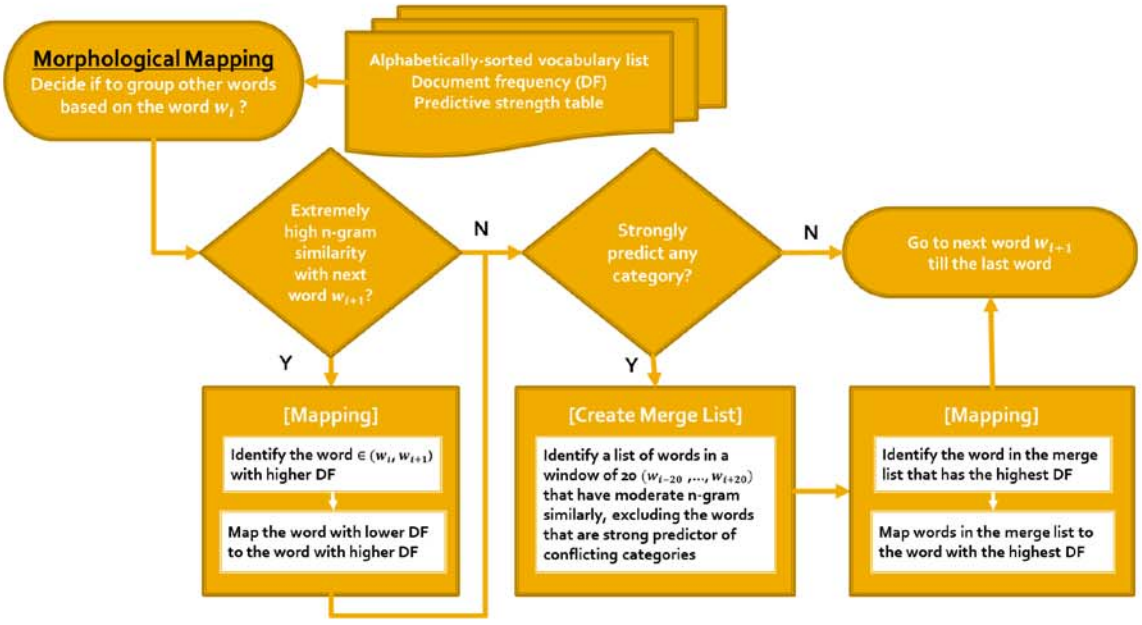

Figure 6.1: Type-M Morphological Mapping Method

As an illustrative example for how Word2Vec differs from Porter's Stemmer and WordNet Lemmatizer in transforming words. Table 6.1 shows a list of words (with the root of "ele") related to electricity in QISU dataset, along with their original DF as well as transformed words and new DF using the proposed Type-M Mapping, Porter's Stemmer, and WordNet Lemmatizer (Python's nltk.stem package). Table 6.1 shows that Type-M Mapping was able to merge most of the listed LFWs to the most frequent word "electric." Porter's Stemmer was able to merge some LFWs to more frequent words, but most were localized grouping. WordNet Lemmatizer was only able to merge two words that were indexed in WordNet, while leaving the rest unchanged.

Table 6.1: Examples of Word Grouping by Type-M Mapping, Porter's Stemmer, and WordNet Lemmatizer

word	DF	Type-M	New_DF	Stem	New_DF	Lemma	New_DF
elec	9	electric	1689		9		9
elecetric	1	electric	1689	elecetr	1		1
elecric	1	electric	1689	elecr	3		1
elecrical	2	electric	1689	elecr	3		2
elecrtic	1	electric	1689	elecrt	1		1
electic	7	electric	1689	elect	14		7
electical	4	electical	4	elect	14		4
electicution	1	electricution	1	electicut	1		1
electirc	3	electric	1689		3		3
electocuted	1	electric	1689	electocut	3		1
electocution	2	electric	1689	electocut	3		2
electri	1	electric	1689		1		1
electric	927	electric	1689	electr	1247		929
electrica	3	electric	1689		3		3
electrical	297	electric	1689	electr	1247		297
electrican	2	electric	1689		2		2
electrici	2	electric	1689		2		2
electricial	5	electric	1689	electrici	6		5
electrician	40		40		40		41
electricians	1	electric	1689	electrician	1	electrician	41
electricity	21	electric	1689	electr	1247		21
electrick	2	electric	1689		2		2
electricla	1	electric	1689		1		1
electrics	2		2	electr	1247	electric	929
electricty	2	electric	1689	electricti	2		2
electriculation	1	electric	1689	electricul	1		1
electricuted	5	electric	1689	electricut	13		5
electricution	8	electric	1689	electricut	13		8
electricy	1	electric	1689	electrici	6		1
electrique	1	electric	1689	electriqu	1		1
electrition	1	electric	1689	electrit	1		1
electriv	1	electric	1689		1		1
electrocute	2	electric	1689	electrocut	371		2
electrocuted	63	electric	1689	electrocut	371		63
electrocution	306	electric	1689	electrocut	371		306
electrode	1	electric	1689	electrod	1		1
electrolysis	1	electric	1689	electrolysi	1		1
electrolyte	1	electric	1689	electrolyt	1		1
electrucuted	1	electric	1689	electrucut	1		1
electuricuted	1	electric	1689	electuricut	1		1

6.2 Experimental Design

The Type-M Mapping method requires the information of predictive category for each word. In machine learning, a trained linear classifier gives a higher coefficient to a feature that has higher discriminatory power. A classifier is forced to put more emphasis on these features with high coefficients when predicting categories of test instances. Thus, the coefficient matrix of a linear classifier serves as an indicator of predictive strength. Features with a high coefficient for a certain category can be said to predict that category. In this study, I used the coefficient matrix of a trained linear classifier to identify the predictive category of words in the feature space (i.e., vocabulary). The predictive category for a given word is defined as the category that has the highest coefficient among all categories for that word.

In this experiment, the coefficient matrix of three classic classifiers were examined, which are as follows:

- MNB
- SVM
- LR

In this experiment, a word was considered as a strong predictor word of a category if the corresponding coefficient was higher than a pre-defined threshold (-10 for MNB and 0.2 for SVM and LR). The threshold was set to a reasonable cut-off where words start to lose the discriminatory power based on the domain knowledge.

The effectiveness of the Type-M Mapping Method was evaluated based on the impact of classification performance of three classic classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). The impact of classification performance was the difference in the macro-averaged F-measure of all categories between mapping and non-grouping. The non-grouping was the standard classification based on the feature space of entire vocabulary.

The effectiveness of three classifiers in identifying predictive categories for words based on their coefficient matrices was related to the level of Type-M Mapping improved the classification performance. The classifier that resulted in a higher positive effect on classification performance was considered to be more effective as an indicator of words' predictive strength for categories.

6.3 Results and Discussion

First, the analysis of variance (ANOVA) was conducted to analyze the differences in the means of impact (i.e., the macro-averaged F-measure across all categories) of Type-M Mapping using the coefficient matrix of three different classifiers (MNB, SVM, LR). In addition to the main factor Coefficient, the analysis also considered the factors (and their interactions) that can potentially influence the impact. These secondary factors were: Category, Classifier (MNB, SVM, LR), and Train-Test Ratio (1:9, 1:1, 9:1). The factor Category was considered as a block factor, which was assumed to have no integration with other factors. The results of ANOVA are listed in Table C.1 in Appendix C. As Table C.1 shows, the factor Coefficient was not statistically significant at an alpha level of 0.05 (F-value = 0.22, P-value= 0.8066). This means that there was no statistically significant difference in the impact of Type-M Mapping among the three coefficient matrices (MNB, SVM, LR). The other two factors, Category and Classifier, were statistically significant. Category was a block factor, and thus will not be discussed. The effect of Classifier will be discussed later in this section.

Table 6.2 lists the impact mean of using the coefficient matrix of three classifiers in three train-test ratio scenarios. Although statistically indifferent, the overall group mean, from highest to lowest, is LR (1.40%), SVM (1.35%), and MNB (1.28%).

Table 6.2: Effect of Type-M Mapping Paired with Coefficient Matrix of MNB, SVM, LR

Coefficient Matrix of Classifier	**1:9**			**1:1**			**9:1**			**Overall**
	MNB	SVM	LR	MNB	SVM	LR	MNB	SVM	LR	
Coef_MNB	3.13%	0.36%	0.64%	3.18%	0.29%	0.29%	3.08%	0.31%	0.25%	1.28%
Coef_SVM	3.28%	0.69%	0.88%	2.93%	0.55%	0.43%	3.00%	0.07%	0.34%	1.35%
Coef_LR	3.51%	0.67%	0.95%	3.21%	0.33%	0.36%	3.14%	0.03%	0.36%	1.40%
Average	3.40%	0.57%	0.82%	3.11%	0.39%	0.36%	3.07%	0.14%	0.32%	1.34%

Next, the effect of Type-M Mapping was examined graphically in Figures 6.2-6.4, with one figure for each train-test ratio scenario. Figure 6.2 shows the effect of the Type-M Mapping under the train-test ratio of 1:9. Applying the Type-M Mapping when the availability of training data is limited, the overall impact on the macro-averaged F-measure, from highest to lowest, was 3.4% for MNB, 0.82% for LR, and 0.57% for SVM. In addition, Type-M Mapping using the coefficient matrices of LR and SVM was found to have slightly better, although statistically insignificant, effects (about +0.3% each) than MNB.

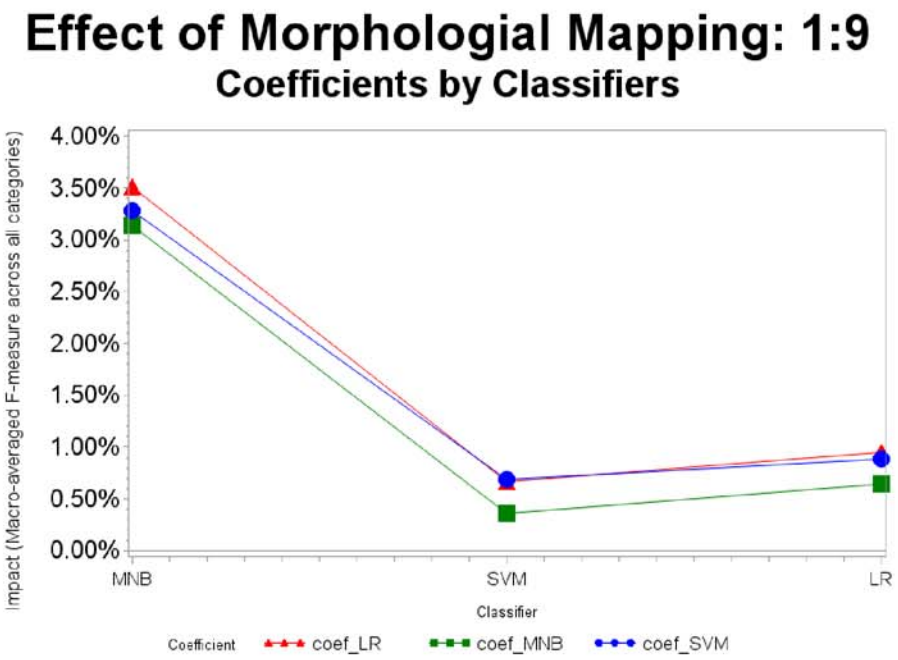

Figure 6.2: Effect of Type-M Mapping: Coefficient by Classifier at Train-Test Ratio of 1:9

Figures 6.3 and 6.4 show the effect of Type-M Mapping in the train-test ratio of 1:1 and 9:1. Similarly to the ratio of 1:9, Type-M Mapping was found to have a much higher positive impact on the classification performance of MNB than SVM or LR. This result confirmed the statistical significance of the factor Classifier in the ANOVA. As for the factor Coefficient, three coefficient matrices were not found to have the difference as obvious as at the ratio of 1:9 in terms of impact. This finding explained the insignificance of the factor Coefficient in the ANOVA test.

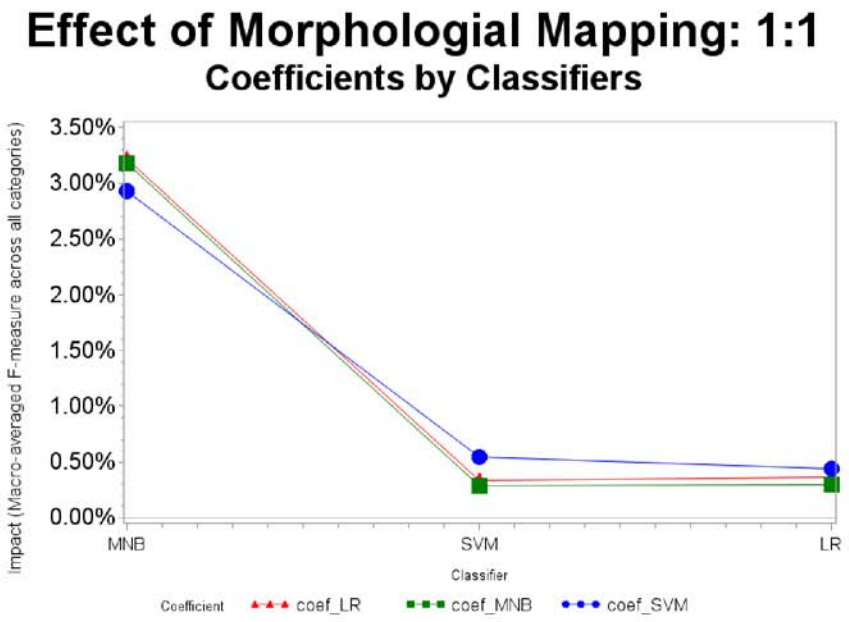

Figure 6.3: Effect of Type-M Mapping: Coefficient by Classifier at Train-Test Ratio of 1:1

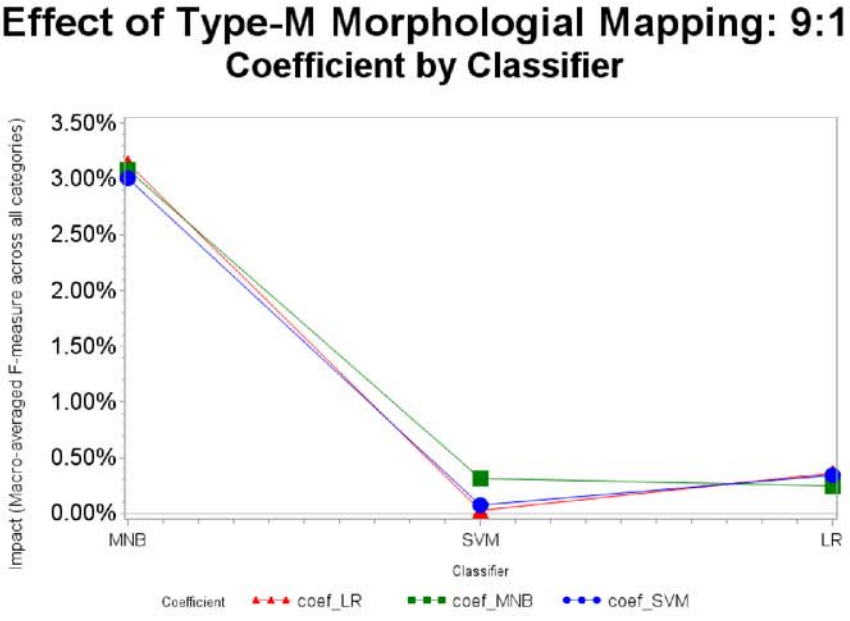

Figure 6.4: Effect of Type-M Mapping: Coefficient by Classifier at Train-Test Ratio of 9:1

As the factor Coefficient was not significant, here I ignored the minimal difference and examined the effect of Type-M Mapping for MNB, SVM, and LR in each train-test ratio. As Figure 6.5 shows, Type-M Mapping was found most effective when the training data is limited (Train-Test Ratio = 1:9) and slightly less effective with sufficient training data (about 0.4% reduction in the improvement).

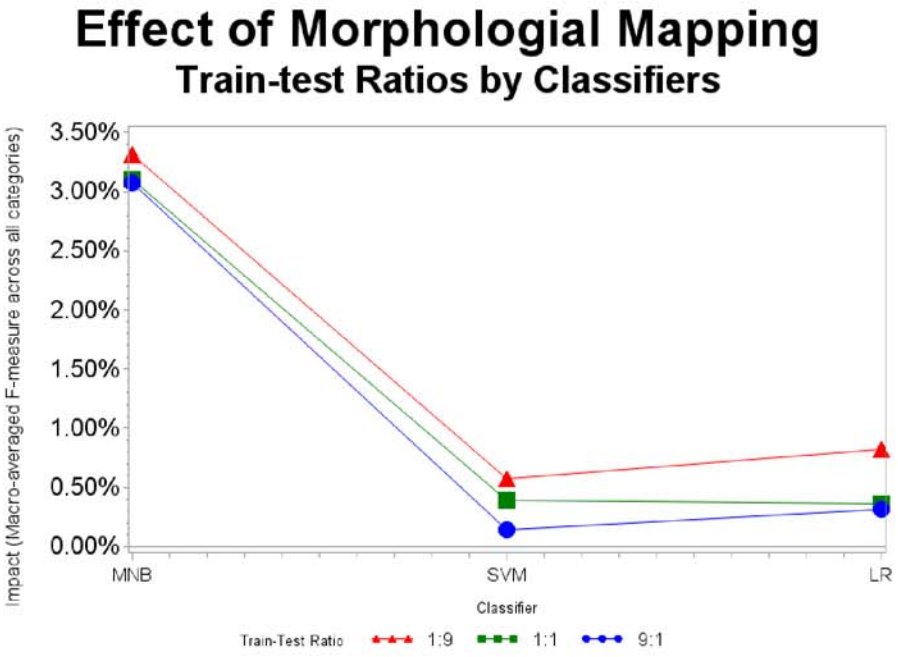

Figure 6.5: Effect of Type-M Mapping: Train-Test Ratio by Classifier

Next, how Type-M Mapping influences classification performance was explored from the perspective of category sizes: large (L), medium (M), and small (S). Table 6.3 summarizes the effect of Type-M Mapping for the three category sizes (L, M, S) and three classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). Overall, Type-M Mapping was found to have the highest positive impact on classification performance of small categories, followed by medium categories. However, Type-M Mapping was found to have a slightly negative impact on large categories.

Table 6.3: Effect of Type-M Mapping on Classification Performance for Different Category Sizes

Category Size	**1:9**			**1:1**			**9:1**			**Overall**
	MNB	SVM	LR	MNB	SVM	LR	MNB	SVM	LR	
L	0.00%	-0.12%	-0.16%	-0.21%	-0.12%	-0.16%	-0.21%	-0.15%	-0.18%	-0.15%
M	3.13%	0.16%	0.15%	2.40%	0.11%	0.14%	1.98%	0.04%	-0.02%	0.90%
S	4.18%	1.33%	2.04%	4.80%	0.92%	0.80%	5.37%	0.34%	0.92%	2.30%
Average	2.44%	0.46%	0.68%	2.33%	0.30%	0.26%	2.38%	0.08%	0.24%	1.02%

Figures 6.6-6.8 graphically show the effect of Type-M Mapping on the classification performance of three category sizes in three train-test ratio scenarios, for MNB, SVM, and LR respectively.

For MNB in Figure 6.6, Type-M Mapping demonstrated the highest positive effect on the classification performance of small categories (5%) and moderate effect on medium categories (2.5%), though insignificantly negative effect on large categories. Although the effect of Train-Test_Ratio was not statistically significant, Type-M Mapping had a higher positive effect on small categories when the train-test ratio is larger while medium categories showed an opposite trend. Type M-Mapping did not impact large categories when the ratio is 1:9 but slightly decreased the performance when the ratio increases to 1:1 or 9:1.

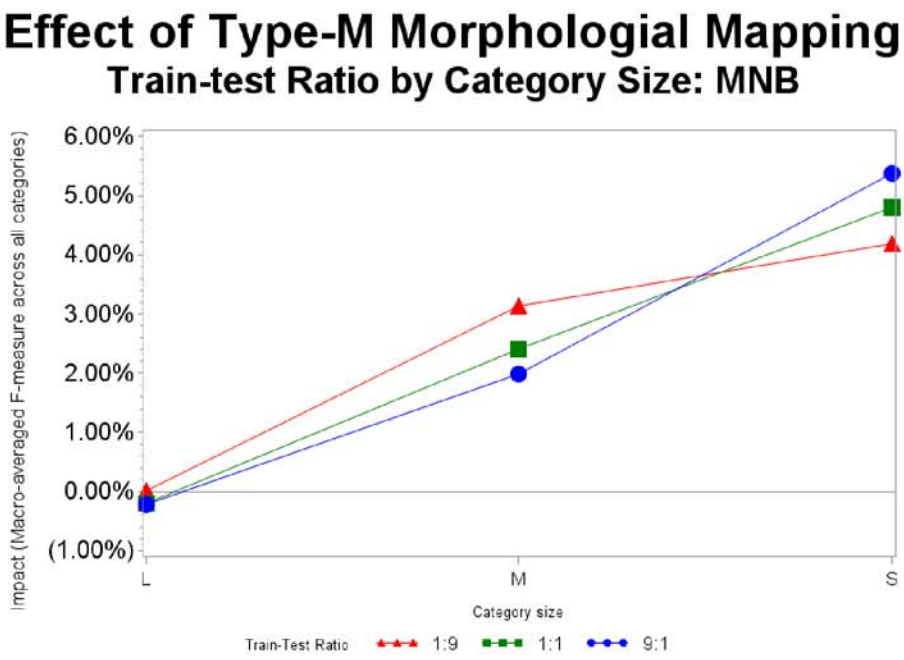

Figure 6.6: Effect of Type-M Mapping: Train-Test Ratio by Category Size for MNB

As for SVM and LR, Type-M Mapping showed a similar pattern of effect on three category sizes, and thus is discussed together. Figures 6.7 and 6.8 show the effect of Type-M Mapping for SVM and LR, respectively. Similar to MNB, the Type-M Mapping also showed the highest positive impact on the classification performance of small categories, though a minor impact on medium categories and slight negative impact on large categories. When the training data is limited (at the train-test ratio of 1:9), Type-M Mapping achieved the highest positive effect on the classification performance of small categories (1.33% for SVM and 2.04% for LR), followed by the ratio 1:1 (0.92% for SVM and 0.8% for LR) and 9:1 (0.34% for SVM and 0.92% for LR). However, the effect of Type-M Mapping was not statistically different among three train-test ratios for all three classifiers.

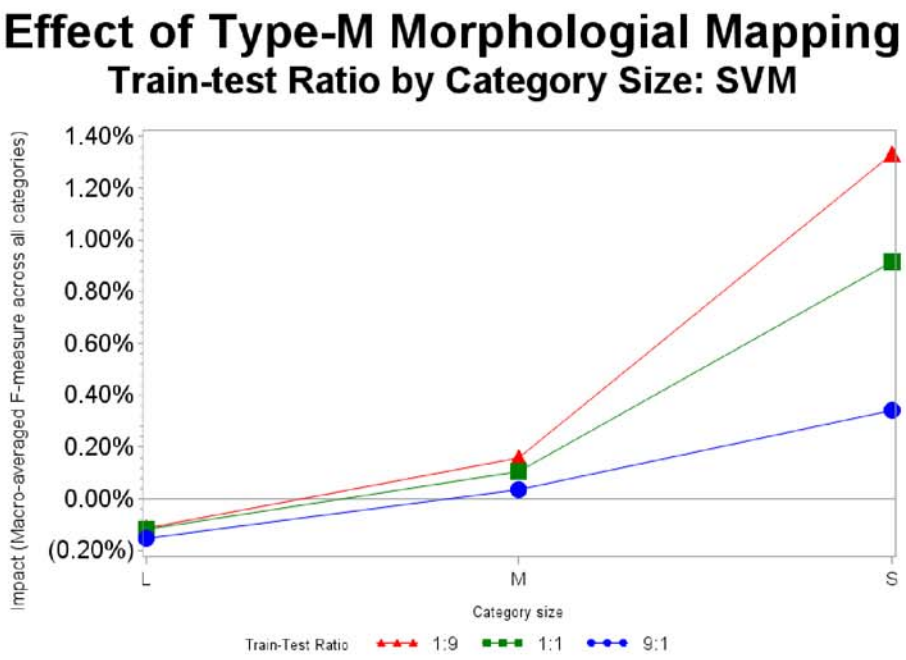

Figure 6.7: Effect of Type-M Mapping: Train-Test Ratio by Category Size for SVM

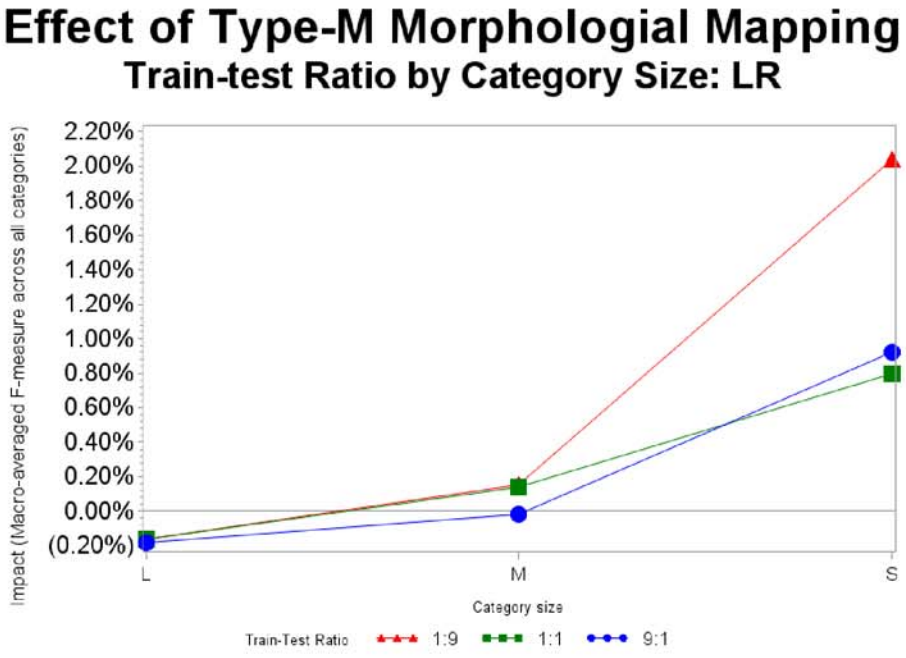

Figure 6.8: Effect of Type-M Mapping: Train-Test Ratio by Category Size for LR

6.4 Summary for Type-M Mapping Method

In Chapter 6, the Type-M Mapping Method was proposed and validated. As a grouping strategy, Type-M Mapping aimed to merge same-hypernym (similar-concept) words that have a similar spelling when they predict a non-conflicting category. This method considered the entire vocabulary (words in train and test sets) and their predictive categories to address the limitations of stemming and lemmatization. Specifically, Type M-Mapping was capable of dealing with misspellings and domain-specific words. Also, Type-M Mapping grouped words in a more conservative but more grounded way because words were merged only when they did not predict conflicting categories. Since the method required the knowledge of predictive categories for each word in the vocabulary list, the coefficient matrices of three classic linear classifiers (MNB, SVM, LR) were examined in terms of how effective these classifiers were in quantifying the predictive strength of words for categories.

The effectiveness of the Type-M Mapping Method was evaluated based on the difference in classification performance (macro-averaged F-measure of all categories) between mapping and non-grouping (i.e., standard classification based on the feature space of entire vocabulary without any grouping). This study compared the effect of Type-M Mapping using three classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). Key findings are highlighted below:

1. Overall, Type-M Mapping was effective in improving the overall classification performance (1.34%), and most effective at the train-test ratio of 1:9 (1.6%), followed by 1:1 (1.29%) and 9:1% (1.18%).

2. Type-M Mapping was found to have a slightly more, although not statistically significant, positive effect when using the coefficient matrix of LR and SVM compared to MNB as an indicator of predictive categories.

3. Type-M Mapping was the most effective in improving the classification performance of MNB (range: 3.07% to 3.4%; mean: 3.19%), followed by LR

(range: 0.32% to 0.82%; mean: 0.5%) and SVM (range: 0.14% to 0.57%; mean: 0.37%).

4. Type-M Mapping was found to have the greatest positive effect on improving classification performance of small categories (+2.3%), followed by medium categories (+0.9%), but a slight negative effect on large categories (-0.15%).

7. TYPE-S SEMANTIC GROUPING METHOD

As noted in Chapter 5, this chapter is dedicated to the Type-S Semantic Grouping Method, the second part of the Type M+S Grouping Method. The main purpose of Type-S Grouping is to group same-hypernym (similar-concept) words with different spelling, supplementing Type-M Mapping which only groups same-hypernym words with similar spelling.

In order to achieve the goal of Type-S Grouping, knowing the meaning of words is required. However, relying on human semantic knowledge and manually grouping words with similar meaning is tedious and impractical. In this chapter, I describe how semantics of words can be derived from text data and measured statistically, and introduce the proposed the Type-S Grouping Method that utilizes statistical semantics and optional human domain knowledge in a form of manual review to identify and group words with discriminatory and similar concepts. With the same goal as Type-M Mapping, Type-S Grouping was developed to utilize rare words and unseen words (i.e., words that only occur in the test set, but not the training set) by proper grouping to reduce the size of vocabulary and feature space, advance the statistical robustness of discriminatory concepts, and ultimately improve the classification performance.

In Section 7.1, I report the feasibility study of statistical semantics for measuring word similarity and identifying same-hypernym words. In Section 7.1.1, I introduce two types of statistical semantics, correlational semantics and distributional semantics, and their measures. In Section 7.1.2, I introduce the proposed Semantic Data Mining Method that aims to utilize the measure of statistical semantics to identify words with the same hypernym, and then compare the effectiveness of the proposed method paired with two different types of statistical se-

mantics measures to identify the superior statistical semantics measure for the use in the following study.

The aim of Section 7.2 is to introduce and examine the so-called Type-S Semantic Grouping Method that utilizes the correlational semantics measured by Word2Vec (the superior statistical semantics measure identified in Section 7.1.2) to group same-hypernym words for the purpose of improving classification performance. Section 7.2.1 discusses the exploratory study conducted for Word2Vec in order to understand its mechanism in quantifying word similarity and ranking similar words in the content of poisoning and allergy (PA) injury data. After gaining a better insight of how Word2Vec operates, in Section 7.2.2 I introduce the Type-S Semantic Grouping Method and discuss its effectiveness in improving the classification performance of the targeted PA categories by grouping words with similar PA-concept. In addition, I also discuss the effect of two grouping strategies (tagging and mapping) and manual review effort (combined effect of threshold and review levels) on the effectiveness of Type-S Grouping. The superior grouping strategy identified is used for the following study.

In Section 7.3, I extend the Semantic Grouping Method proposed in Section 7.2.2 to the classification of all categories, not limited to only PA categories. To reduce the effort of manually selecting seedwords as discriminatory concepts for grouping, the possibility of automating this process is explored. As the literature has shown the capability of feature selection methods for prioritizing features, I report the effectiveness of classic feature selection methods in identifying discriminatory concepts for semantic grouping and improving the classification performance. As tagging is identified as a superior strategy for semantic grouping in Section 7.2.2, I also report the effectiveness of the automated Type-S Tagging Method paired with the feature selection method on improving classification performance.

7.1 Statistical Semantics: Correlational and Distributional Semantics

7.1.1 Introduction

In natural language processing, there are two types of word relation or word similarity. The 1-st order word relation, also called syntagmatic relation, concerns the position. Words are syntagmatically related if they co-occur in a text more frequently than by chance. Syntagmatical word-pairs tend to neighbor each other and can be in any semantic relationships such as synonyms, meronyms, antonyms, and words that are functionally related or frequently associated. The measure of syntagmatic relation can be derived from the word co-occurrence information. Examples of classic measures include pointwise mutual information (PMI), t-test, chi-square test, and log-likelihood ratio. PMI was considered to be superior to other 1-st order relation measures in syntactic and semantic tasks tasks (L. Han, Finin, McNamee, Joshi, & Yesha, 2013; Pecina, 2005). PMI measures the likelihood that two words tend to co-occur versus occurring alone, which is analogic to the correlation in statistics which measures the degree that two variables tend to co-increase / decrease. In this sense, the statistical similarity of words measured by the 1-st order measures can be referred to as "correlational similarity" (Han et. al., 2013).

On the other hand, the 2-nd order relation, or paradigmatic relation, concerns the substitution. Paradigmatic word-pairs tend to have similar neighbors and are often substitutable for one another in a specific context, thus they are likely to be synonyms or antonyms or share a hypernym. The method to quantify the 2-nd order word relation between two words can be measured by the similarity of their neighboring words or context, which is often represented by the cosine of two context vectors. The 2-nd relation is also called "distributional similarity" since it assumes that words with a similar context tend to share a similar meaning (i.e., distributional hypothesis). The 2-nd order approaches to measure the 2-nd word relation or distributional semantics can also be called "distributional seman-

tics models (DSMs)". The DSMs can be classified into count-based (unsupervised) models and predict-based (supervised) models according to the nature of methods for deriving the representation of the context vectors. The unsupervised / count-based DSMs are based on the transformation or reweighting of the original word co-occurrence matrix, often involving singular value decomposition to reduce the dimensionality (e.g. latent semantic analysis). On the other hand, the supervised / predict(ive)-based DSMs utilize the neural probabilistic language model to predict the context given a target word (or vice versa) by framing text data as a supervised task without involving any manual annotation. One of the predict-based DSMs, Word2Vec (Mikolov et al., 2013), developed by Google in 2013, can be considered as the most promising example of a predict-based DSM due to its possibility of capturing the linguistic features of human language, both syntactically and semantically. As a predict-based model, Word2Vec was also found to outperform count-based models on syntactic and semantic tasks (Baroni et al., 2014).

Table 7.1: Statistical Semantics Similarly: Correlational Similarity and Distributional Similarity

	1st-order / Correlational Similarity	**2nd-order / Distributional Similarity**
Definition	"Position": two words co-occur more frequently than by chance and they are likely to be the neighbor of each other	"Substitution": two words tend to have similar neighbors and they are often substitutable for one another
Example	Bank-trust company; car-wheel; hot-cold; knife-cut; pencil-paper	Doctor-nurse; apple-orange (relatively tighter semantic relationship)
Measurement Basis	Occurrence and co-occurrence frequencies	Context vectors under the "distributional hypothesis"
Representative Method	Pointwise mutual information (PMI) $$PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)}$$ $$\approx \log \frac{Count(w_1, w_2) \times N}{Count(w_1) \times Count(w_2)}$$	Cosine of the context vectors that can be derived from: • Count-based DSMs: Transformation or reweighting of original matrix (Example: Latent semantic analysis) • Predict-based DSMs: predict the context given a target word or vice versa based on self-annotated text data (Example: Googles Word2Vec).

PMI and Word2Vec seem to be the best representative for the measure of correlational and distributional similarity. As there has been no study on the effectiveness of correlational similarity and distributional similarity in identifying words with similar concepts in the content of injury surveillance data, I intend to provide an answer with an experiment of utilizing correlational and distributional similarity measures to identify the agent of injury (i.e., name of drug) from injury

narratives of drug poisoning and allergy cases. This is done by identifying the word in a narrative that is most similar to the pre-selected comparing seed-word ("drug," for example). In the following sections, I discuss the experimental design and results, followed by the hypothesis testing regarding the effectiveness of correlational and distributional similarity. The superior measure identified in this experiment was used for the following experiments of utilizing semantic grouping for improving classification performance.

7.1.2 Evaluation of Statistical Semantics in Identifying Same-hypernym Words

This experiment serves as a pilot study to identify an effective similarity measure that is capable of finding synonyms or words with a similar hypernym or concept statistically in an injury dataset. The results were published at the AHFE 2016 conference (Huang, Nanda, Lehto, & Vallmuur, 2016).

In this experiment, I examined the effectiveness of the correlational similarity measured by PMI and distributional similarity measured by Word2Vec in a task of identifying same-hypernym words in injury text data. Hypothesis 5 was tested by comparing the performance of PMI and Word2Vec in the task of identifying the injury agent (i.e., the drug name) from drug-related injury narratives.

The proposed semi-supervised semantic data mining method integrates human review as verification into the machine prediction process. The performance was evaluated based on two criteria: accuracy (in terms of the proportion of test cases with a correctly identified injury agent) and manual review effort (in terms of the proportion of vocabulary required for review).

(i) **Semantic data mining method**

A semantic data mining method was proposed, which utilizes the statistical semantics measure to identify the agent of injury (i.e., the name of drug) from injury narratives of drug poisoning and allergy cases, specifically by finding the word that has the highest similarity score against with a pre-selected

seedword (e.g. "drug") based on a similarity measure (PMI or Word2Vec). This method allows the human input in a form of manual review to incorporate into the machine prediction process. The procedure of the semantic data mining method is listed below and Figure 7.1:

1. Select the following variables:
    a. Similarity measure: PMI or Word2Vec
    b. Threshold for a word to be considered as the drug name candidate
    c. Seedword that each word in the narrative compares against
    d. Stopping criteria (maximum accuracy, number of runs for manual reviews, number of words to review)
2. Prepare word similarity score for each possible word with the seed-word
3. [Prediction by machine] For each word that is not listed as the non-drug word in each injury narrative:
    a. calculate the similarity between the word and the seedword
    b. designate the word whose similarity score is the largest and beyond the threshold as the drug name candidate
4. [Verification by human reviewer] Review a unique list of drug name candidates
    a. Classify them into drug or non-drug names and create or update the drug wordlist and non-drug wordlist
5. Repeat Steps 3 and 4 until the stopping criteria is met

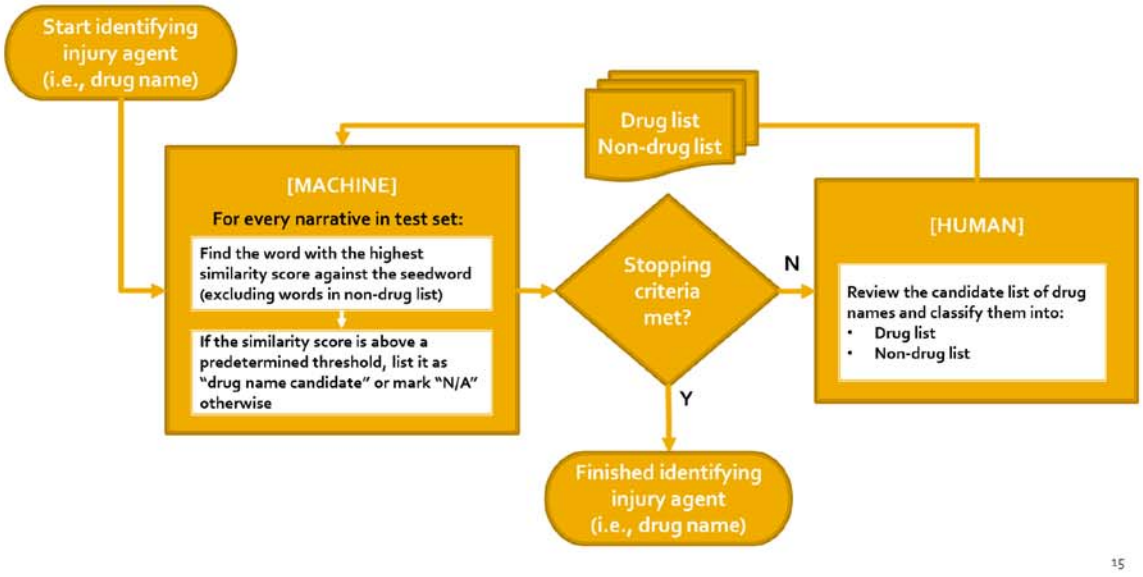

Figure 7.1: Semantic Data Mining Method to Identify Agent of Drug Poisoning and Allergy Related Injury

(ii) **Experimental design**

In this experiment, three factors, as independent variables, were considered:

- Two similarity measures
  - Correlational similarity: PMI
  - Distributional similarity: Word2Vec
- Ten threshold levels — for a word to be considered as the drug name candidate
- Two seedwords:
  - general-concept seedword: "drug"
  - specific-concept seedword: "panadol" (which is a brand of painkiller in Australia)

Two performance measures, as dependent variables, were evaluated for each combination of independent variables.

1.) Manual review effort, measured by the proportion of vocabulary (e.g. 10% of manual review effort = 10% of vocabulary)

2.) Accuracy, measured by the proportion of test cases with correctly identified injury agent

The dataset for evaluation was the injury narratives that were coded as 17 (Poisoning due to drug or medicinal substance) for their external cause, which accounted for about 1% of the QISU dataset. This sample was comprised of 5581 narrative texts of drug poisoning and allergy injury, which had 65 thousand word occurrences and 5747 unique words. The narratives were all lowercased after removing non-English characters and stopwords except for prepositions.

As for the preparation of statistic semantic similarity, PMI utilizes the word co-occurrence information while Word2Vec relies on the context vectors that were learned from the corpus by the predict-based DSM. Given the word frequency and co-occurrence matrix, the correlational similarity, measured by PMI, for words w_1 and w_2 is calculated based on the following equation:

$$PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)} \approx \log \frac{Count(w_1, w_2) \times N}{Count(w_1) \times Count(w_2)} \tag{7.1}$$

where $P(w_1, w_2)$ and $Count(w_1, w_2)$ are the probability and the document counts that w_1 and w_2 co-occur, $P(w_1)$ and $P(w_2)$ are the individual probabilities of w_1 and w_2, $Count(w_1)$ and $Count(w_2)$ are the counts of documents where w_1 and w_2 is present, N is the total number of word occurrences in the corpus being analyzed.

The Word2Vec model was trained on the QISU injury corpus where non-English characters are removed and remaining words are lower-cased. The original feature space with 47 thousand unique words was downsized to 300 newly generated features, using the default setting of Word2Vec in Python.

The stopwords were retained because the Word2Vec was found to perform better without removing any words in the pilot study. Compared to excluding stopwords, Word2Vec trained on the entire vocabulary including stopwords achieves the same accuracy with less manual review effort. The distributional similarity score, measured by Word2Vec, between words w_1 and w_2 is calculated by the cosine similarity of the two context vectors in 300-dimensional space.

$$Word2Vec\ similarity = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|} \tag{7.2}$$

where v_1 and v_2 are the newly-derived context vector.

(iii) **Results and discussion**

First, the performance of the correlational similarity measured by PMI was examined. Figure 7.2 shows the relation between accuracy and manual review effort in percentage under 10 different levels of threshold. In Figure 7.2, the left plot focuses on the general-concept seedword "drug" and the right plot focuses on the specific-concept seedword "panadol." The initial accuracy for using PMI to identify drug names without involving any manual review was around 15% with the seedword "drug" and less than 10% with "panadol." The accuracy increased as the manual review effort increased. For example, with the seedword "drug", 10% of manual review effort significantly increased the accuracy by 45%, from 15% to 60%, after the first run of manual review. However, with the seedword "panadol", 10% of manual review effort improved the accuracy by around 20%, half of the improvement with "drug". Both graphs also show that a lower threshold was more likely to push the limit of maximum accuracy (i.e., last point of each line for each threshold level) higher.

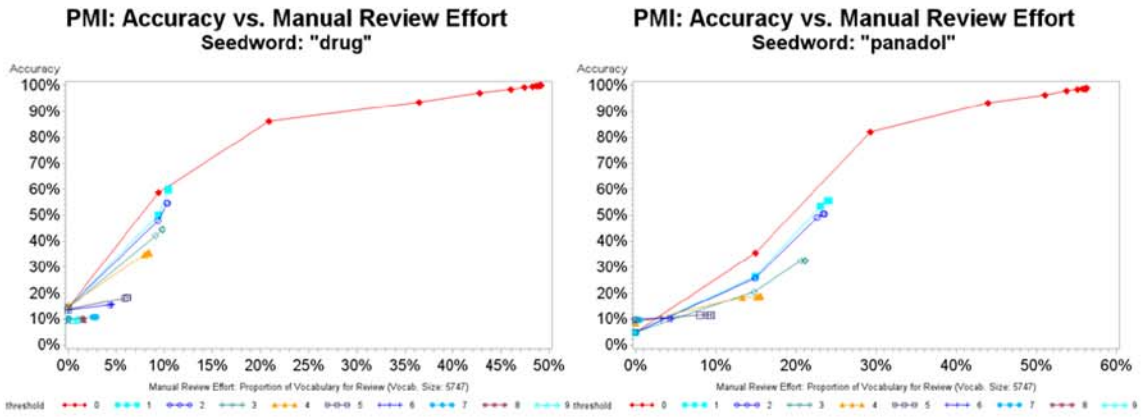

Figure 7.2: Correlational similarity measured by PMI: Accuracy vs. Manual Review Effort (Left: with seedword *drug*; Right: with seedword *panadol*)

Two similar plots for the distributional similarity measured by Word2Vec are shown in Figure 7.3. It can be observed that the initial accuracy without any manual review is much higher than using PMI: 30% with the seedword "drug" and 65% with seedword "panadol", compared to 15% and 10% for PMI. 10% of manual review effort was able to achieve an accuracy of 80% for Word2Vec with either the seedword "drug" or "panadol."

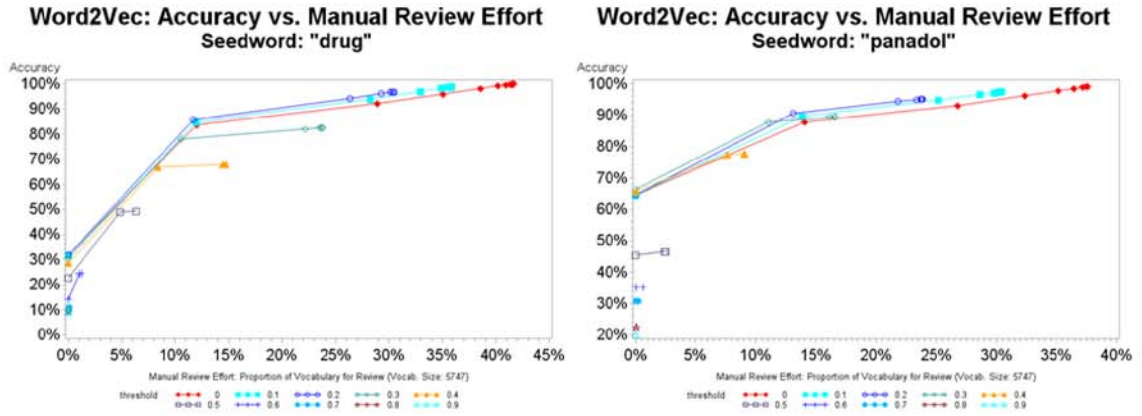

Figure 7.3: Distributional similarity measured by Word2Vec: Accuracy vs. Manual Review Effort (Left: with seedword *drug*; Right: with seedword *panadol*)

Next I discuss the maximum accuracy that PMI and Word2Vec were able to achieve at ten different threshold levels along with the required effort of manual review. Figure 7.4 shows such relationship for PMI with a bar-line chart on the left plot and a line chart on the right plot. In the bar-line chart of Figure 7.4, bars (right: *panadol*; left: *drug*) indicate the proportion of vocabulary for manual review effort, represented by the left Y-axis while lines indicate the maximum accuracy, represented by the right Y-axis. X-axis represents ten different levels of thresholds. The results suggest that a lower threshold tends to have a higher maximum accuracy and to involve more effort of manual review.

The line chart on the right of Figure 7.4 can be used to examine which seedword works better for the PMI; the left line is for "drug" and right line is for "panadol." The line chart directly compares the maximum accuracy and manual review effort. The results indicate that PMI worked better with the seedword "drug" because this combination achieved a higher accuracy with less manual review effort.

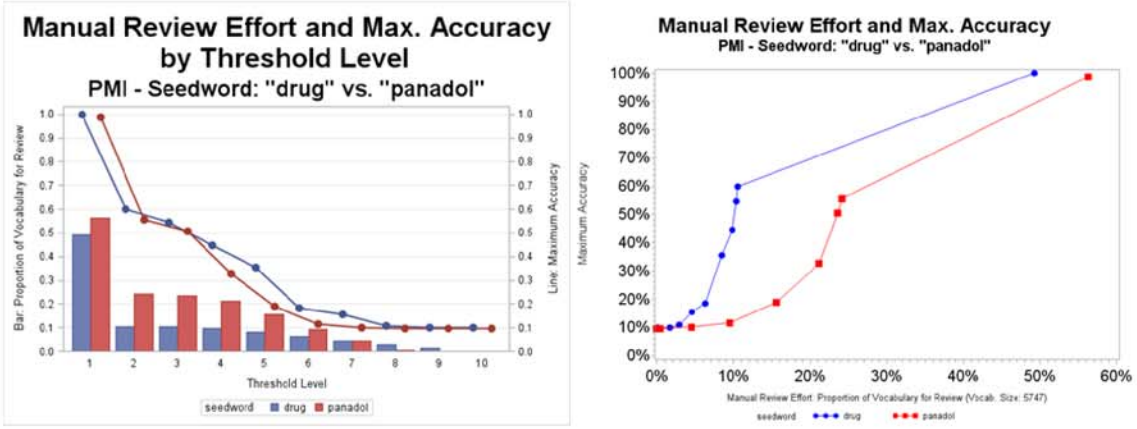

Figure 7.4: Manual Review Effort vs. Maximum Accuracy: PMI — seedword *drug* vs. *panado*

Figure 7.5 shows the graphs of bar-line chart (left) and line chart (right) for Word2Vec, organized similarly to Figure 7.4. The line chart (right: *drug*; left: *panadol*) indicates that the specific-concept seedword "panadol" was more effective for Word2Vec.

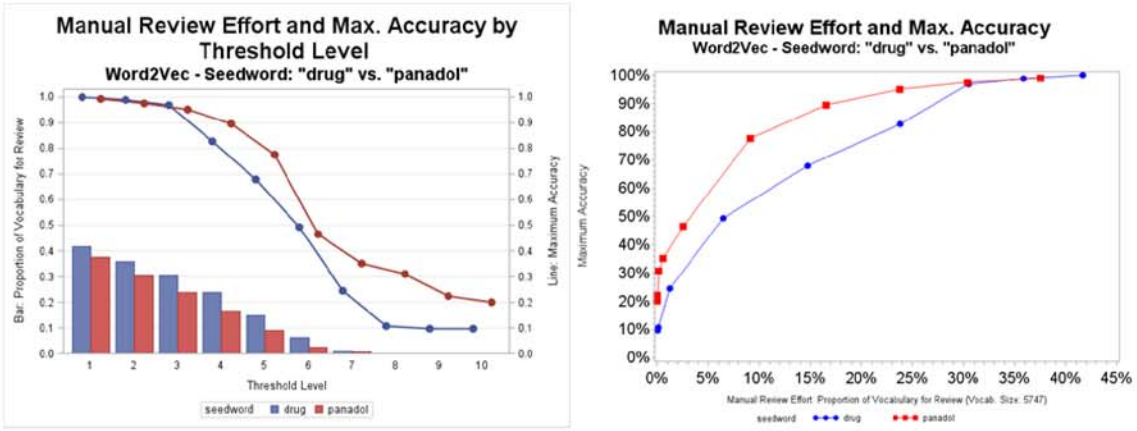

Figure 7.5: Manual Review Effort vs. Maximum Accuracy: Word2Vec — seedword *drug* vs. *panadol*

In the following, similar bar-line charts and line charts are used to compare the performance between PMI and Word2Vec using different seedwords. Figure 7.6 is for the seedword "drug" and Figure 7.7 is for the seedword "panadol."

Both Figures 7.6 and 7.7 show the consistent results that Word2Vec outperformed PMI, either with the seedword "drug" or "panadol", because it required less manual review effort to achieve the same level of accuracy.

By comparing the line charts (right: PMI; left: Word2Vec) of Figures 7.6 and 7.7, it can be observed that the gap between two lines was more significant when the seedword "panadol" was used. The result indicated that PMI required more manual review effort to achieve the same accuracy with the specific-concept seedword "panadol" than the general-concept seedword

"drug". This was because Word2Vec performed better with the seedword "panadol" while PMI performed worse with it than "drug."

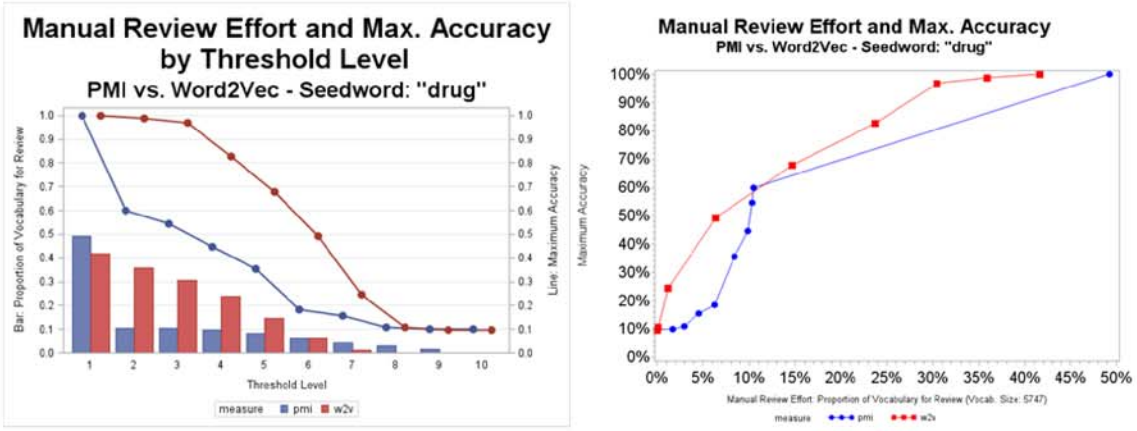

Figure 7.6: Manual Review Effort vs. Maximum Accuracy: Seedword *drug* — PMI vs. Word2Vec

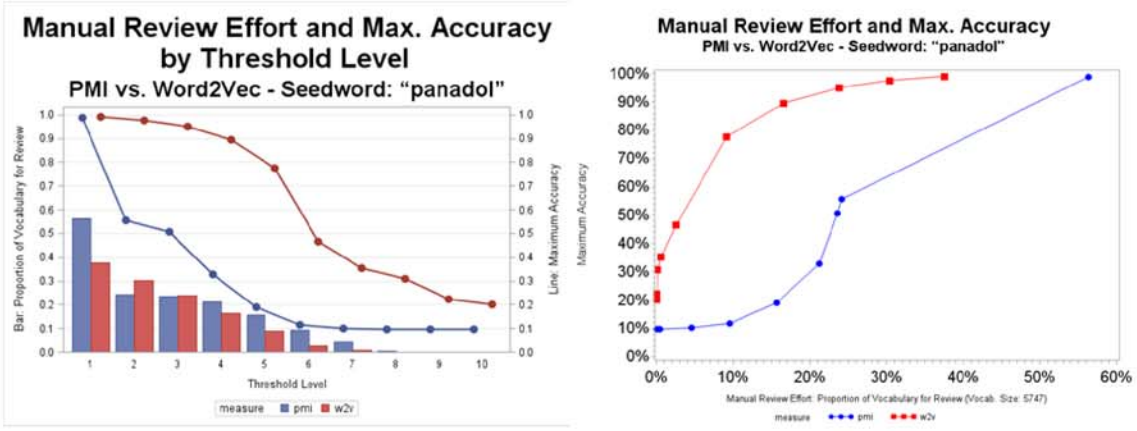

Figure 7.7: Manual Review Effort vs. Maximum Accuracy: Seedword *panadol* — PMI vs. Word2Vec

To sum up the findings, regardless of which seedword was used, Word2Vec outperformed PMI in the task of identifying words with similar concept ("drug" in this experiment). Word2Vec performed better with the concept-

specific seedword "panadol" while PMI did not, which made their performance difference more significant when the concept-specific seedword "panadol" was used.

The results falsified Null Hypothesis 5 and accepted Alternative Hypothesis 5: *"Distributional similarity measured by Word2Vec is more effective than the correlational similarity measured by PMI in identifying words shared with the same hypernym in injury surveillance data."*

The explanation for these results is provided in the following. Correlational similarity measured by PMI is based on co-occurrence. Two pairing words need to co-occur frequently to have a higher PMI similarity, but drugs were not mentioned frequently with other drugs or the exact word "drug" in the same narrative. The specific-concept seedword "panadol" had a lower frequency and thus had a lower chance to be mentioned with other drugs or "drug." This explains why PMI was less effective in finding the words with the drug concept in the QISU dataset.

On the other hand, the distributional similarity measured by Word2Vec is based on the similarity of context. The general-concept seedword "drug" is a more common word. The word "drug" was mentioned in narratives of various injury circumstances, and thus it had a context vector with a wider range of features. As "drug" tended to associate with many words, it became relatively less associated with the words that indicate a drug. In contrast, the concept-specific seedword "panadol" had a smaller set of active context features (such as routes of entry, symptoms, reason, etc.), which were often shared by other drugs. This, in turn, made the specific-concept seedword "panadol" a good candidate for comparing with other words and for identifying words with the similar concept of drug.

7.2 Semantic Grouping with Word2Vec

In Section 7.1, the Word2Vec-based semantic method was found effective in identifying words with similar concept, specifically the concept of drug in the injury narratives of drug poisoning or allergy "PA" related cases.

In this section, I extend the method to other PA related cases, not limited to only cases associated with drug, and test if the grouping of similar-concept words can improve the classification performance of the related categories. In the QISU surveillance data, PA related cases were originally coded 17 (poisoning due to drug or medicinal substance) and 18 (poisoning due to other or unspecified substance) for external cause of injury. Those cases were further classified into sub-categories, called "PA categories" for simplicity, according to their type of injury agent: drug (PA_DRUG), chemical (PA_CHEMICAL), alcohol (PA_ALCOHOL), drug and alcohol (PA_DRUGALCO), food (PA_FOOD), plant (PA_PLANT), and other or unspecified (PA_OTHERS).

Since the original semantic data mining method in Section 7.1 focuses on a single type of injury agent, the selection of seedword involves only one concept "drug". In Section 7.2, multiple types of injury agents and concepts are targeted, including: drug, chemical, alcohol, food, and plant. These concepts are essential for classifying PA categories, which are called "PA-concepts" for simplicity in the following. The seedwords that represent these concepts need to be properly selected so as to group words with a similar concept or injury agent for the purpose of potential classification improvement.

In Section 7.2.1, similar logic for selecting seedwords as the previous experiment is followed. I selected seedwords with the generic-concept and specific-concept for each PA category, excluding PA_DRUGALCO and PA_OTHERS. I looked into the top 100 words most similar to each selected seedword and examined their association based on the semantic and domain knowledge. This exploratory study not only provided a foundation for selecting seedwords for the extended semantic

grouping method in Section 7.2.2, but also helped us understand the mechanism of Word2Vec in terms of measuring statistical semantics and identifying similar words.

In Section 7.2.2, the semantic grouping method for PA concepts is first introduced in the part (a). PA concepts are essential to proper classification of PA categories. This grouping method aims to identify and group the words with a similar concept to pre-selected seedwords, which represent the PA concept (identified in Section 7.2.1). The ultimate goal is to improve the classification performance of PA categories. Next, I elaborate the experimental design in the part (b) for evaluating the impact of semantic grouping along with three factors that can potentially influence the performance (i.e., grouping strategy, threshold, review level). Lastly, the results are discussed in the part (c) and summarized in the part (d).

7.2.1 Exploratory Study for Word2Vec

As an unsupervised learning technique, Word2Vec is a measure of distributional similarity that relies on context vectors to quantify statistical semantics. Two words with a high Word2Vec similarity score often share similar context (of documents) and are often associated with each other semantically or syntactically. Common associations between word-pairs include synonym ("unhappy" and "sad"), antonyms ("true" and "false"), hyponym – hypernym (as "car" to "vehicle"), and meronyms – holonym (as "wheel" to "car"), or they are just frequently mentioned together ("cooking" and "kitchen"), or functionally related ("would" and "like").

The objective of this study is to utilize the Word2Vec similarity measure for identifying words that have the same hypernym. In Section 7.2.1, Word2Vec showed its capability to identify words that indicate drug with the general-concept seedwords "drug" or specific-concept seedword "panadol." To better understand the mechanism of Word2Vec, I conducted an exploratory study for Word2Vec and

examined how similar words are ranked. I focused on two PA-concepts (drug and chemical), which were essential to classify the categories of PA_DRUG and PA_CHEMICAL. The results also provided a basis to select seedwords for the Semantic Grouping Method introduced in Section 7.3, where I report an experiment that tested if the grouping of words with the same PA-concept can improve the classification performance of PA categories.

Similar logic in Section 7.2.1 was followed for selecting seedwords. For each PA-concept, at least one general-concept and specific-concept seedwords were selected for investigation. Table 7.2 lists the two PA-concepts (drug and chemical) and associated seedwords to examine. In the seedword list, the first word is a general-concept seedword, which has the same form as its associated PA-concept. The following words are specific-concept seedwords, which are hyponyms of the associated PA-concept.

Table 7.2: Exploratory Study for Word2Vec — Selected Concepts and Words

PA-concept	**Seedwords**
DRUG	*drug, panadol, advil, cocaine, diazepam, antibiotic, zyrtec*
CHEMICAL	*chemical, detergent, shampoo, paint*

After training a Word2Vec model on the QISU dataset, the top 100 most similar words to each selected seedword were identified. Note that Word2Vec may yield a slightly different result on different runs due to a random seed involved in optimization.

The first step of the exploratory study was, for each seedword, to check if the word in the top 100 list was a hyponym of the associated PA-concept. If not, I identified the role of this word by examining the injury narratives within which it was mentioned. Note that if the word was a typo, the role was assigned based on its correct spelling. Additional note was made if the word was a typo or part of

multi-word phrases. The two PA-concepts, drug and chemical, are discussed respectively for each selected seedword, followed by a summary of general findings for their associated PA-concept.

PA-concept: DRUG

According to the Health Services of Northern Territory Government (n.d.), "drug" is defined as "any chemical substance, natural or synthetic, that changes a person's mental state and that may be used repeatedly by a person for that effect." A drug can be associated with one or more drug classes. A drug class is a set of drugs that have something in common, either having similar chemical structures or the same mechanism of action, and/or being used to treat the same disease (PubMed Health, 2015). Common drug classes include analgesic, antibiotic, sedative, antidepressant, recreational, and antihistamine (allergy medication). In this section, in addition to the general-concept seedword "drug," several specific-concept seedwords are selected to investigate. These specific-concept seedwords are either the name, brand, or class of a common drug from several different drug classes, which are listed in Table 7.3.

Table 7.3: Drug-related Seedwords and Their Drug Classes for Word2Vec Exploratory Study

Drug Class	**Seedword(s)**
analgesic	*panadol, advil*
recreational	*cocaine*
sedative	*diazepam*
antibiotic	*antibiotic*
allergy medication	*zyrtec*

Panadol and Advil are analgesic or pain relief products. Both can treat pain and reduce fever, but Panadol has a weaker effect than Advil (Skwarecki, 2015). Panadol is a brand name of an analgesic called Acetaminophen, which treats minor aches and pains. On the other hand, Advil is a brand name of a nonsteroidal anti-inflammatory analgesic called Ibuprofen, which treats fever and mild to severe pain. Panadol seems to be a more common painkiller than Advil in Australia. Since Panadol and Advil are the same class of drug but have different popularity (i.e., DF in the injury dataset), it is of interest to compare their Word2Vec results. Cocaine is a recreational drug, which has a purpose different from therapeutic drugs. Diazepam is a sedative, which can treat anxiety, alcohol withdrawal symptoms, muscle spasms, and seizures. Antibiotics are used to treat or prevent bacterial infections. Zyrtec is a brand for an allergy medication.

The top 100 most similar words (called "top words" for simplicity) to each seedword were examined. The common roles and their definitions are listed in Table 7.4. If the word was identified as drug, it was further classified into an appropriate drug class.

Table 7.4: Roles of Most Similar Words to Drug-related Seedwords

Role	Description
same-hypernym	Hyponym of drugs, including drug class (such as "analgesic"), specific name (such as "ibuprofen"), brand (such as "advil"), or a word that is not any of the above but indicates drug (such as "polypharmacy")
symp	Symptom
route	Route-of-entry
unit	Measurement unit or a noun that implies amount or quantity
chemical	Chemical
alcohol	Alcohol
food	Food
other (noise)	Other: a word that has no semantic meaning or discriminatory power. Usually is an extremely low-frequency word.
other (descriptive)	Other: a word describing the injury activity such as how, who, when, where, what involved, treatment, etc.
other (bx as treatment)	Other: symptom or disease that can be treated by antibiotics

PA-concept: DRUG – Seedword "drug"

First, the top 100 most similar words ("top words" for simplicity) to the general-concept seedword "drug" was examined. As Table 7.5 shows, the accuracy of identifying hyponyms of drug using the seedword "drug" was 60%. The next major portion of top words was noisy words (17%), considered as an artifact of Word2Vec due to its favor towards rare events. 7% of top words were symptoms, including overdose, intoxication, poisoning, od, and misspellings of these related words. In addition, 6% of top words were describing words. For example, intentional, and its misspellings intensional and intentionaltook, implying that drug-related poisoning and allergy could often be intentional. Consistent with the QISU dataset, it was also observed that many injury narratives had the phrases such as "inten-

tional overdose" or "intentional od." Another 5% of top words were associated with the measurement unit, which was also an inevitable artifact of Word2Vec as drugs were often mentioned along with quantities and measurement units such as amount, quantities, and tablets. Furthermore, 2% of top words were part of multi-word drugs, such as "illicit" in "illicit drug" and "prescription" in "prescription medication." The rest 3% of top words were related to alcohol, chemical, and route-of-entry due to the similar context (poisoning related) that they shared with "drug." Note that 44% of top words similar to "drug" were low-frequency words ($DF <= 3$), which suggested that Word2Vec tended to give a higher weight to rare events.

Table 7.5: Role of Top 100 Most Similar Words to *drug*

Role	%	Words Similar to "drug"
same-hypernym	60%	*polypharmacy, ectasy, valpro, anphetamines, ecstasy, xanex, polypharm, immovane, guarana, drugs, heroin, cocaine, lsd, ativan, ecstacy, temaz, stillnox, fantasy, lorazepam, alprazolam, paroxetine, seroquel, ducene, xanax, zanax, vallium, phernergen, xdoxylamine, codalgin, luvox, prothiadine, viagra, duramine, olazepam, antidepressant, coveram, lexapro, tranqualizers, oxazepam, xydep, marijuana, ginseng, thyroxin, temaze, cones, valium, doxylamine, dimicron, temazepam, norvasc, nitrazepam, chlorambucic, respiradol, cytotoxic, propanalol, artane, cafergot, temaxe, cobantrim, tamapam*
other (noise)	17%	*breathiliser, onwards, unuse, asccompanied, admited, relating, taylor, inactives, nisha, uticaria, absynth, htoh, nencur, nemacur, tabletand, tablest, premix*
symp	7%	*poisoning, intoxication, overdose, intoxicationbal, od, poisoningod, poisoningalleged*
other (descriptive)	6%	*intentionaltook, intentional, intensional, schoolie, user, bibqps*
unit	5%	*quantities, quantity, excessive, cask, amounts*
other (partial)	2%	*illicit, prescription*
alcohol	1%	*vodka*
chemical	1%	*carbon*
route	1%	*consumption*

Next, injury narratives that contained the word "drug" in the QISU dataset were investigated to understand the typical context of "drug." Table 7.6 lists the categories of the documents that contained "drug" and their count and percentage statistics. It was observed that the word "drug" is often associated with drug plus alcohol poisoning (PA_DRUGALCO; 49.5%), drug poisoning (PA_DRUG; 20.4%), or any injuries due to the use of drug (and alcohol) such as alleged assault

(STRUCKCOLLISION; 10.3%), self-harm and suicidal ideation (CUTTING; 5.8%), and fall (FALL; 5.1%).

Table 7.6: Category Distribution for Documents that Contain the Word *drug*

Category	Count of Documents that Contain "drug"	%
ANIMAL	10	1.1%
BICYCLE	3	0.3%
C289	20	2.3%
C289FBI	1	0.1%
CHOKING	3	0.3%
CUTTING	**51**	**5.8%**
FALL	**45**	**5.1%**
FIREFLAME	2	0.2%
HORSE	1	0.1%
HOTOBJ	1	0.1%
MACHINERY	1	0.1%
MOTORCYCLE	10	1.1%
MOTORVEHICLE	8	0.9%
PA_ALCOHOL	2	0.2%
PA_CHEMICAL	14	1.6%
PA_DRUG	**179**	**20.4%**
PA_DRUGALCO	**435**	**49.5%**
PA_FOOD	1	0.1%
PEDESTRIAN	1	0.1%
STRUCKCOLLISION	**90**	**10.3%**
Grand Total (DF)	878	

PA-concept: DRUG – Seedword "panadol"

Panadol is a common brand of painkiller in Australia. Given the seedword "panadol", the accuracy of identifying hyponyms of drug was 77%. In addition to the noisy words as an artifact of Word2Vec (13%), other top words include: unit

(5%), part of multi-word drug names (such as "NUROFEN ADVANCE", "PAIN RELIEF", "PANADOL RAPID"; 3%), and descriptive word (as "plus" in "overdose Panadol plus alcohol"; 1%). The results are listed in Table 7.7 below:

Table 7.7: Roles of Top 100 Most Similar Words to *panadol*

Role	%	Words Similar to "panadol"
same-hypernym	78	*nurofen, paracetamol, pandol, neurofen, paracetemol, ibuprofen, advil, nurfen, painstop, loratadine, pnandol, ibrufen, panamax, voltaren, nurefen, zavance, tramal, xnurofen, panafen, endone, brufen, paracetamolat, mersyndol, nuofen, telfast, panadiene, digesic, panadine, paracetomol, zyrtec, claryntine, codapane, naprosyn, clarytine, pandadeine, polarimine, anlalgesia, prodeine, fent, zirtek, pandadol, metaclopramide, panadeeine, panadeine, fenac, padadol, indocid, bruffen, pandeine, codeine, phenergen, celebrex, elixer, panadols, mdma, neufron, capadex, ketorolac, nsaids, codapaine, claratyne, panadolx, valiums, topomax, phenergan, analagesia, aspro, herron, nurophen, codeie, mobic, phernergan, nuerofen, fortex, panadole, voltarin, antihistamine, oesto*
other (noise)	13	*yesdterday, som, nilknown, ambulancewith, aleast, clinginess, tartan, risible, miinimal, triagemhistory, onrouts, pentrharna, migrane*
unit	5	*grm, gram, gm, mlls, foils*
other (partial)	3	*advance, releif, rapid*
other (descriptive)	1	*plus*

Panadol is a common drug used to relieve pain for a variety of injuries. As Table 7.8 suggests, 10.8% of documents that contained "panadol" (584 out of 5410) were associated with drug poisoning (PA_DRUG). "Panadol" was frequently mentioned in the context where patients were prescribed Panadol as pain relief for various types of injuries including: fall (Fall; 39.3%), struck by or struck against

(STRUCKCOLLISION; 24%), overexertion (C289; 10%), and burnt by hot objects (HOTOBJ; 3.4%). It is also worthy to mention that Panadol was the most commonly overdosed drug in the QISU dataset, accounting for 10.5% of total 5578 PA_DRUG cases. The reason could be due to its high accessibility to the public.

Table 7.8: Category Distribution for Documents that Contain the Word *panadol*

Category	Count of Documents that Contain "panadol"	%
ANIMAL	110	2.0%
BICYCLE	122	2.3%
C289	**542**	**10.0%**
C289EYE	24	0.4%
C289FBEAR	10	0.2%
C289FBI	5	0.1%
C289FBNOSE	1	0.0%
CHOKING	1	0.0%
CUTTING	119	2.2%
ELECTRICITY	2	0.0%
FALL	**2126**	**39.3%**
FIREFLAME	17	0.3%
HORSE	38	0.7%
HOTCOLDCOND	9	0.2%
HOTOBJ	**183**	**3.4%**
MACHINERY	38	0.7%
MOTORCYCLE	75	1.4%
MOTORVEHICLE	37	0.7%
OTHERTRANSPORT	14	0.3%
PA_CHEMICAL	6	0.1%
PA_DRUG	**584**	**10.8%**
PA_DRUGALCO	38	0.7%
PA_FOOD	2	0.0%
PA_PLANT	1	0.0%
PEDESTRIAN	6	0.1%
STRUCKCOLLISION	**1300**	**24.0%**
Grand Total (DF)	5410	

In addition, 78% of top 100 most similar words to "panadol" are same-hypernym words that indicate drugs. Table 7.9 shows the further classification of these 78 drugs according to their drug classes. 59% of drugs identified had the same class, analgesic, as Panadol. Other commonly associated drug classes include anti-inflammatory (18%) and allergy medication (17%). Anti-Inflammatory is a drug used to reduce inflammation or swelling, which remedy pain as well (U.S. Food and Drug Administration, 2016). On the other hand, allergy medication such as antihistamine can relieve discomfort caused by allergy by blocking one type of receptor for histamine, which is a chemical that is responsible for many of the signs and symptoms of allergic reactions (Ogbru, 2015). Thus, anti-Inflammatory and allergy medication are similar to analgesics due to the shared purpose of easing pain and physical discomfort. Word2Vec was able to capture such association.

Table 7.9: Drug Classes of Top 100 Most Similar Drugs to *panadol*

Drug Class	**Count of Drugs**	**%**
Analgesic	46	59%
Anti-Inflammatory	14	18%
Allergy Medication	13	17%
Anticonvulsant	1	1%
Recreational	1	1%
Sedative	1	1%
Osteoporosis treatment	1	1%
Gut motility stimulator	1	1%
Total	78	

PA-concept: DRUG – Seedword "advil"

Another drug-related seedword "advil" is also an analgesic like Panadol. Advil is a brand name of a nonsteroidal anti-inflammatory analgesic called Ibuprofen, which treats fever and mild to severe pain. Although both are analgesics, Advil

was mentioned far less frequently than Panadol in the QISU dataset ("advil" has a document frequency of 58 compared to 5410 for"panadol"). It is of interest to examine if Word2Vec would derive similar results for using these two similar-concept, but different-frequency seedwords. Table 7.10 shows the associations for the top 100 most similar words to "advil." 90% of top words were drugs. Despite its low document frequency, "advil" has a higher accuracy (90%) of identifying words that indicate drug than"panadol" (78%).

Table 7.10: Drug Classes of Top 100 Most Similar Drugs to *advil*

Role	%	Words Similar to "advil"
same-hypernym	90	*zavance, voltaren, endone, panamax, capadex, telfast, panafen, mobic, paracetemol, celebrex, tramal, mersyndol, loratadine, ibrufen, oxynorm, voltarin, tramadol, indocid, fenac, phenerghan, ibuprofen, digesic, diclofenac, pandol, codapane, restavit, prodeine, naprosyn, brufren, codine, claratyne, lexopro, paracetomol, codiene, codeine, topomax, zyprexia, aspro, phenergen, bruffen, panadiene, valiums, murelax, codapaine, paracetamolat, escitalopram, imovane, epilum, zyrtec, prestiq, nuofen, diazapam, herron, nsaids, chlonazepam, pandadeine, neurofen, xnurofen, sudafed, zolpidem, thermoslim, avanza, brufen, zirtek, valpam, histamine, neulactil, pandeine, elixer, paracetamol, aspalgin, neufron, antihypertensive, ibruprofen, atenolol, metaclopramide, rabieprazole, oesto, pandiene, osteo, naproxen, indomethacin, stillnox, seroquil, nurofin, nurophen, kapadex, claryntine, ramipril, codalgin*
unit	5	*gram, grm, gm, omg, extra*
other (partial)	3	*rapid, rel, advance*
other (noise)	2	*foils, vomtied*

Table 7.11 shows that the contexts of "advil" were often associated with injuries of, from highest to lowest frequency, fall (Fall; 38%), drug poisoning (PA_DRUG;

22%), overexertion (C289; 21%), and being struck (STRUCKCOLLISION; 14%). The results in Table 7.11 demonstrated a category distribution similar to"panadol" in Table 7.8. Both Advil and Panadol were most likely to occur in the three largest categories of external causes (i.e., FALL, STRUCKCOLLISION, and C289) and drug poisoning (PA_DRUG).

Table 7.11: Category Distribution for Documents that Contain the Word *advil*

Category	**Count of Documents that Contain "advil"**	**%**
ANIMAL	1	2%
C289	**12**	**21%**
CUTTING	1	2%
FALL	**22**	**38%**
MOTORCYCLE	1	2%
PA_DRUG	**13**	**22%**
STRUCKCOLLISION	**8**	**14%**
Grand Total (DF)	58	

In addition to the similar contexts, Table 7.12 shows that "advil" had the same top three associated drug classes as"panadol", which are: analgesic (46%), anti-inflammatory (13%), and allergy medication (12%). In addition to these three shared drug classes, "advil" (11) were associated with more irrelevant drug classes than"panadol" (6). The tendency to have noisy results could be due to the low occurrence of "advil" in the QISU dataset, which explained the fundamental flaw of statistical semantics to deal with rarity.

Table 7.12: Drug Classes of Top 100 Most Similar Drugs to *advil*

Drug Class	Count of Drugs	%
Analgesic	41	46%
Anti-Inflammatory	12	13%
Allergy Medication	11	12%
Sedative	6	7%
Anticonvulsant	3	3%
Antidepressant	3	3%
Antihypertensive	3	3%
Antipsychotic	3	3%
Proton-Pump inhibitor	2	2%
Osteoporosis treatment	2	2%
Decongestant	1	1%
Gut motility stimulator	1	1%
Hypnotic	1	1%
Weight-loss medication	1	1%
Total	90	

PA-concept: DRUG – Seedword "cocaine"

Another drug-related seedword "cocaine" was examined. Cocaine is a recreational drug, which is also known as a "street drug" that is taken for nonmedical purpose (Farlex Partner Medical Dictionary, 2012). Common recreational drugs include cocaine, marijuana, heroin, amphetamines, methamphetamine, ecstasy, and LSD. Recreational drugs are considered powerfully addictive stimulants that cause short-term health effects such as extreme happiness and energy, mental alertness, and hypersensitivity to sight, sound, and touch (National Institute on Drug Abuse, 1969). The word "cocaine" had only seven occurrences in the QISU dataset. It would be interesting to explore how Word2Vec would operate given such a rare feature. As Table 7.13 suggests, 69% of top words (similar) to "cocaine" were ac-

curately identified as drugs. The rest 31% of top words were associated with measurement unit (13%), alcohol (8%), noise (3%), symptoms (2%), describing word (2%), part of multi-word drugs (2%), and route-of-entry (1%).

Table 7.13: Roles of Top 100 Most Similar Words to *cocaine*

Role	%	Words Similar to "cocaine"
same-hypernym	69%	*ecstasy, ectasy, ecstacy, cones, zanax, paroxetine, valpro, stillnox, heroin, lsd, xanax, duramine, temaz, marajuana, doxylamine, lexapro, alodorm, oxazepam, risperdal, restavit, methadone, serepax, alprazolam, zolpidem, metformin, luvox, seroquel, fantasy, nitrazepam, aropax, marijuana, thc, endep, temazapam, serequil, digoxin, antidepressants, diamicron, antabuse, zoloft, valium, benzos, cipramil, epilim, oxycontin, codalgin, zyprexa, zyoban, seroquil, antidepressant, ssri, tamapam, mersyndal, baclofen, cannabis, lasix, valuim, duromine, mirtazapine, stilnox, guarana, tramadol, ativan, diclofenac, invega, psychotic, atenolol, tegretol, valproate*
unit	13%	*cask, pills, amounts, casks, quantities, pill, premix, volume, qty, quantity, omg, grams, cans*
alcohol	8%	*malibu, shots, vodka, bourbon, scotch, alchol, whisky, cruisers*
other (noise)	3%	i*ngingestion, lager, admited*
symp	2%	*paranoid, overdose*s
other (descriptive)	2%	*intensional, induced*
other (partial)	2%	*herbal, prescription*
route	1%	*smoked*

The word "cocaine" only occurred in seven documents in the QISU dataset. All seven occurrences were poisoning and allergy related. As Table 7.14 indicates, three cases were drug poisoning (PA_DRUG; 43%) and four were drug and alcohol poisoning (PA_DRUGALCO; 57%).

Table 7.14: Category Distribution for Documents that Contain the Word *cocaine*

Category	Count of Documents that Contain "cocaine"	%
PA_DRUG	3	43%
PA_DRUGALCO	4	57%
Grand Total (DF)	7	

As Table 7.15 shows, 12% of drugs identified as the top 100 most similar words to "cocaine" were recreational, the same drug class as "cocaine." Other top associated drug classes include sedative (22%), antidepressant (17%), analgesic (10%), antipsychotic (10%), and anticonvulsant (6%). These drug classes are all considered to be psychoactive drugs along with Cocaine and other recreational drugs. Psychoactive drugs are used to describe "any chemical substance that affects mood, perception or consciousness as a result of changes in the functioning of the nervous system" (Northern Territory Government Health Services, n.d). In general, psychoactive drugs can be broadly classified into three categories: "Depressants" that slow down the central nervous system to dull the senses, induce sleep, relieve pain or anxiety, or treat mood disorder or seizure, for example: morphine and heroin; "Stimulants" that excite the nervous system to stimulate the mind, for example: amphetamines, cocaine, and MDMA (Ecstasy); "Hallucinogens" that distort how things are perceived, for example: LSD and cannabis (Baofu, 2011; World Health Organization, 2004; Northern Territory Government Health Services, n.d). Psychoactive drugs can be used for therapeutic or recreational purposes. For therapeutic purpose, psychoactive drugs are prescribed to "reduce or eliminate the suffering caused by psychological conditions such as anxiety, insomnia, depression, psychosis, and bipolar affective disorder" (Addiction Prevention Center, 2008). The drug classes associated with therapeutic, psychoactive drugs include sedative, hypnotics, analgesics, antidepressants, antipsychotics, and anticonvulsant (Psychology Encyclopedia, 2016). In addition, psychoactive drugs

can evoke feelings of euphoria by acting in the brain in different ways. Thus, such types of psychoactive drugs, also known as recreational drugs, are often overdosed for personal pleasure or satisfaction rather than the therapeutic purpose. Table 7.15 indicates that although the first top drug class was not recreational (i.e., the drug class of Cocaine), these top six drug classes associated with "cocaine" are also related as they all belong to psychoactive drugs.

Table 7.15: Drug Classes of Top 100 Most Similar Drugs to *cocaine*

Drug Class	Count of Drugs	%
Sedative	15	22%
Antidepressant	12	17%
Recreational	8	12%
Analgesic	7	10%
Antipsychotic	7	10%
Anticonvulsant	4	6%
Weight-loss medication	3	4%
Alcoholism medication	2	3%
Allergy Medication	2	3%
Oral antidiabetic	2	3%
Antihypertensive	1	1%
Diuretic	1	1%
Heart medication	1	1%
Hypnotic	1	1%
Muscle relaxant	1	1%
obsessive-compulsive disorder	1	1%
Orexigenic	1	1%
Total	69	

PA-concept: DRUG – Seedword "diazepam"

Diazepam is a sedative that can treat anxiety disorder or alcohol withdrawal symptoms. As Table 7.16 shows, 94% of top 100 most similar words to "diazepam" are drugs, with the remaining 4% for measurement unit and 2% for parts of multi-word drugs.

Table 7.16: Roles of Top 100 Most Similar Words to *diazepam*

Role	%	Words Similar to "diazepam"
same-hypernym	94%	*valium, seroquel, alprazolam, temazepam, zoloft, endep, xanax, efexor, zanax, sertraline, oxazepam, temaz, serequel, avanza, esipram, largactil, escitalopram, pristiq, mirtazapine, zyprexa, fluoxetine, neulactil, tramadol, stillnox, nitrazepam, tegretol, effexor, clonazepam, paroxetine, zyprexia, lexapro, serepax, polypharmacy, temazapam, lovan, olanzapine, stilnox, mirtazon, setraline, amitriptyline, risperidone, frusemide, dothiepin, diazapam, restavit, epilim, duromine, fluvoxamine, diamicron, venlafaxine, kalma, zopiclone, dexamphetamines, temaze, citalopram, vallum, oxycontin, atenolol, oxycodone, pseudephi, serapax, ativan, endone, lorazepam, logician, codral, champix, periactin, hydrochloride, lamictal, plaquenil, zocor, lamotrigine, metformin, cipramil, gliclazide, omeprazole, antidepressant, epilum, risperdal, dimicron, serequil, invega, pizotifen, parecetamol, codalgin, diaz, norvasc, antidepressants, zolpidem, prozac, dextrameph, maleate, chlorpheniramine*
unit	4%	*tablest, tabs, omg, gms*
other (partial)	2%	*prescription, sr*

Diazepam is a medium frequency word in the QISU dataset, and mostly associated with poisoning of drug (PA_DRUG; 61%) and drug plus alcohol (PA_DURGALCO; 20.2%) as shown in Table 7.17. Due to the fact that "diazepam" is mostly associated with poisoning and allergy related context, such "pure" context of "diazepam" could explain its high accuracy of identifying same-hypernym words.

Table 7.17: Category Distribution for Documents that Contain the Word *diazepam*

Words Similar to "diazepam"	**Count of Documents that Contain "diazepam"**	**%**
BICYCLE	1	0.5%
C289	14	6.6%
CUTTING	1	0.5%
FALL	9	4.2%
MOTORVEHICLE	1	0.5%
OTHERTRANSPORT	1	0.5%
PA_ALCOHOL	1	0.5%
PA_CHEMICAL	2	0.9%
PA_DRUG	**130**	**61.0%**
PA_DRUGALCO	**43**	**20.2%**
STRUCKCOLLISION	10	4.7%
Grand Total (DF)	213	

As Table 7.18 shows, the drug classes mostly associated with "diazepam" include antidepressant (24%), sedative (21%), antipsychotic (12%), analgesic (10%), anticonvulsant (7%), and recreational (3%). Although sedative, the same class as Diazepam, was not ranked first, these top associated drug classes resemble one another because they are all psychoactive drugs which can "result in alternation in perception, mood, consciousness, cognition, or behavior" (CTI Reviews, 2016).

Table 7.18: Drug Classes of Top 100 Most Similar Drugs to *diazepam*

Drug Class	Count of Drugs	%
Antidepressant	23	24%
Sedative	20	21%
Antipsychotic	11	12%
Analgesic	9	10%
Anticonvulsant	7	7%
Allergy Medication	4	4%
Oral antidiabetic	4	4%
Recreational	3	3%
Antihypertensive	2	2%
Cold and flu	1	1%
Decongestant	1	1%
Diuretic	1	1%
General	1	1%
Hypnotic	1	1%
Immunosuppressant	1	1%
Obsessive-compulsive disorder	1	1%
Proton-Pump inhibitor	1	1%
Smoking cessation aid	1	1%
Statin	1	1%
Weight-loss medication	1	1%
Total	94	

PA-concept: DRUG – Seedword "antibiotic"

Another drug-related seedword examined was "antibiotic," which was another major drug class different from those discussed above (analgesic, recreational, sedative). As Table 7.19 shows, 63% of top words to "antibiotic" were drugs, followed by descriptive words (13%), symptoms or diseases that require antibiotics as treatment (6%), food and chemicals that can cause allergy (5%), symptom (2%),

route-of-entry (2%), and noisy word that happened to share similar context with "antibiotic" (1%).

Table 7.19: Roles of Top 100 Most Similar Words to *antibiotic*

Role	%	Words Similar to "antibiotic"
same-hypernym	63%	*benzos, adrenalin, adrenaline, antihistamine, antihistamines, claratyne, claryntine, epipen, histamine, histamines, phenergan, promethazine, redipred, telfast, zirtec, zyrtec, abx, amethocaine, aspro, astrix, capadex, celebrex, chesty, inflammatory, voltarin, chloro, amoxicillin, amoxil, amoxycillin, amoxyl, antibiotics, augmentin, bactrim, ceclor, cephalexin, chloromycetin, chlorsig, clindamycin, erythromycin, flucloxacillin, keflex, rulide, sylvazine, tonsillitis, baclofen, antidepressents, arapax, cymbalta, dothep, nilstat, mobic, salazopyrin, respiridone, charcoal, trifeme, contrast, cortisone, hydrocortisone, prednisolone, prednisone, steroids, amphetamines, decongestant*
other (descriptive)	13%	*daily, bronchoscopy, pred, prescribed, script, specialist, therapy, viral, sees, bloods, poabs, levels*
other ("abx as a treatment")	6%	*circumcision, gastro, rhinorrhoea, tonsilitis, varicella, uti*
other (partial)	6%	*neb, ointment, duo, vaccine, puffer, worm*
unit	3%	*doses, micrograms, dosage*
food	3%	*dairy, wheat, paste*
chemical	2%	*lotion, menthol*
symp	2%	*hallucinating, hyperactivity*
route	2%	*ingesting, orally*
other (noise)	1%	*thismorning*

As Table 7.20 shows, the contexts of "antibiotic" were often associated with drug poisoning (PA_DRUG; 37%), followed by wound infection caused by animal

bite (ANIMAL; 14%), struck or alleged assault (STRUCKCOLLISION; 14%), fall (FALL; 9%), and others.

Table 7.20: Category Distribution for Documents that Contain the Word *antibiotic*

Category	Count of Documents that Contain "antibiotic"	%
ANIMAL	**5**	**14%**
C289	**3**	**9%**
C289EYE	2	6%
CUTTING	2	6%
FALL	**3**	**9%**
HOTOBJ	1	3%
PA_DRUG	**13**	**37%**
PA_FOOD	1	3%
STRUCKCOLLISION	**5**	**14%**
Grand Total (DF)	35	

Table 7.21 lists the associated drug classes with the top drugs that are identified as similar to "antibiotic." 29% of drugs identified were the same class, Antibiotics. Other top drug classes included allergy medication (24%), analgesic (14%), and steroid (8%). Analgesics are related to antibiotics because they are often used together for different effects: antibiotics only help to clear the infection while analgesics such as Paracetamol relieve the pain caused by infection (NHS Choices, 2015). In addition, both antibiotics and steroid can treat sinus infection, which is often associated with the inflammation of allergic rhinitis (MD Medical Reference, 2016). In this sense, antibiotics, analgesic, steroid, and allergy medication are related drug classes. However, the rest of associated drug classes (such as antidepressant) seem to be less relevant to antibiotics, which could be due to the low occurrence of the word "antibiotic" in the QISU dataset.

Table 7.21: Drug Classes of Top 100 Most Similar Drugs to *antibiotic*

Drug Class	Count of Drugs	%
Antibiotics	18	29%
Allergy Medication	15	24%
Analgesic	9	14%
Steroid	5	8%
Antidepressant	4	6%
Anti-Inflammatory	2	3%
Alcoholism medication	1	2%
Anesthetic	1	2%
Anticonvulsant	1	2%
Antifungal medication	1	2%
Antipsychotic	1	2%
Decongestant	1	2%
Drug overdose treatment	1	2%
Obsessive-compulsive disorder	1	2%
Oral contraceptive	1	2%
Oral contrast for CT	1	2%
Total	63	

PA-concept: DRUG – Seedword "zyrtec"

The last drug-related seedword was "zyrtec." Zyrtec is a brand name of antihistamine or allergy medication called Cetirizine. As Table 7.22 shows, 88% of the top words similar to "zyrtec" were drugs. Other association included quantity (7%), part of multi-word drugs (2%), descriptive words (2%), and chemical (1%).

Table 7.22: Roles of Top 100 Most Similar Words to *zyrtec*

Role	%	Words Similar to "zyrtec"
same-hypernym	88%	*polaramine, claratyne, paracetomol, diclofenac, telfast, capadex, antihistamine, phenergan, amoxicillin, codiene, antidepressant, ceclor, prednisolone, amoxycillin, aspro, codeine, claryntine, prednisone, endone, stemetil, paracetemol, voltarin, erythromycin, buscopan, sudafed, mersyndol, homeopathic, phenergen, digesic, polarimine, demazin, panafen, histamine, tramal, panamax, celebrex, claritine, tramadol, advil, stemitil, indocid, doxycycline, zirtec, naproxen, zirtek, mobic, respiridone, temaze, diazapam, clopidogrel, metformin, serequil, hydrocortisone, contraceptive, benadryl, prozac, epipen, elixer, oxycodone, inflammatory, decongestant, codalgin, phenerghan, loratadine, cymbalta, oxazepam, zyprexa, fluvoxamine, bactrim, epilum, lamictal, tegratol, amoxil, promethazine, stillnox, topomax, duramine, diamicron, fenac, dymadon, cephalexin, lorazapam, antibiotic, codine, thyroxine, norvasc, naprosyn, imovane,*
unit	7%	*mils, doses, gram, omg, mcg, dosage, sip*
other (partial)	2%	*syrup, rapid*
other (descriptive)	2%	*pred, potentially*
other (noise)	1%	*bate*

The word "zyrtec" is not a high-frequency words in the QISU dataset. Table 7.23 shows the contexts associated with "zyrtec." It can be observed that the injury narratives that contain "zyrtec" were mostly associated with poisoning and allergy caused by food (PA_FOOD; 28%), drug (PA_DRUG; 23%), or others or unspecified (PA_OTHERS; 23%), and animal or insect bite (ANIMAL; 23%).

Table 7.23: Category Distribution for Documents that Contain the Word *zyrtec*

Category	Count of Documents that Contain "zyrtec"	%
ANIMAL	**9**	**23%**
C289	1	3%
MACHINERY	1	3%
PA_DRUG	**9**	**23%**
PA_FOOD	**11**	**28%**
PA_OTHERS	**9**	**23%**
Grand Total (DF)	40	

Table 7.24 shows top drug classes associated with Zyrtec. 20% of top drugs similar to "zyrtec" were the same class, allergy medication. Other associated drug classes such as analgesic (31%) and antibiotics (10%) were related to allergy medication because analgesics are contained in some allergy medication to relieve the pain (News Medical, 2007) and antibiotics are often used to treat the sinus infection caused by allergic rhinitis (MD Medical Reference, 2016). In addition to these top 3 associated drug classes, the remaining drug classes seem less relevant. The results, again, could be considered as the limitation of Word2Vec to deal with low frequency features.

Table 7.24: Drug Classes of Top 100 Most Similar Drugs to *zyrtec*

Drug Class	Count of Drugs	%
Analgesic	27	31%
Allergy Medication	18	20%
Antibiotics	9	10%
Antipsychotic	5	6%
Sedative	5	6%
Anti-Inflammatory	4	5%
Anticonvulsant	3	3%
Antidepressant	3	3%
Steroid	3	3%
Oral antidiabetic	2	2%
Stimulant	2	2%
Antihypertensive	1	1%
Blood thinners	1	1%
Decongestant	1	1%
Hypnotic	1	1%
Hypothyroidism Medication	1	1%
Oral contraceptive	1	1%
Weight-loss medication	1	1%
Total	88	

Summary of Word2Vec Exploratory Study for Drug-related Seedword

In this section, I examined several drug-related seedwords, including one general-concept seedword "drug" and six specific-concept seedwords. These specific-concept seedwords can be a drug name (such as "cocaine", "diazepam"), drug brand (such as"panadol", "advil", "zyrtec"), or drug class (such as "antibiotic"). These seedwords have different document frequencies (DF) in the QISU corpus, from the lowest DF of 7 for "cocaine" to 5410 for "panadol." Due to the inherent differences in their purposes of use, these seedwords had different occurrence patterns and con-

texts in the QISU dataset, which in turn influenced how Word2Vec modeled their relationship and quantified the similarity between them and other words. The top 100 most similar seedwords for each seedword were examined, and further classified according to their roles in the QISU dataset. The roles included: same-hypernym (i.e., drug), symptoms, route-of-entry, measurement unit, substances that could cause poisoning or allergy (such as chemical, alcohol, and food), and others that frequently shared similar context (such as parts of multi-word drugs, describing words for injury incidents, diseases or symptoms that can be treated by antibiotics, and noisy words that do not have semantic meaning or discriminatory power). Table 7.25 list the distribution of roles and DF for the top 100 most similar words to each drug-related seedword. The injury narratives associated with the seedword were also explored and grouped by external cause in order to identify typical contexts for each seedword.

Table 7.25: Distribution of Roles and Document Frequency for Top 100 Most Similar Words to Drug-related Seedwords

	Drug-related seedword (Document Frequency)						
Role	"drug" (878)	"panadol" (5410)	"advil" (58)	"cocaine" (7)	"diazepam" (213)	"antibiotic" (35)	"zyrtec" (40)
same -hypernym (accuracy)	60%	78%	90%	69%	94%	63%	88%
symp	7%			2%		2%	
route	1%			1%		2%	
unit	5%	5%	5%	13%	4%	3%	7%
chemical	1%					2%	
alcohol	1%			8%			
food						3%	
other (noise)	17%	13%	2%	3%			1%
other (partial)	2%	3%	3%	2%	2%	6%	2%
other (descriptive)	6%	1%		2%		13%	2%
other ("abx as a treatment")						6%	
associated documents related to poisoning and allergy	71.8%	11.7%	22%	100%	82.6%	40%	74%

As Table 7.25 shows, the general-concept seedword "drug" had accuracy of 60% while the specific-concept seedwords were able to achieve at least 63% accuracy (with "antibiotic"), and up to 94% (with "diazepam"), to identify words that indicate drug. This result was consistent with the finding in Section 7.1 that, for distributional similarity that relies on the context similarity, using a specific-concept seedword tended to be more effective in identifying same-hypernym words. This was because a specific-concept seedword was more likely to have similar context (poisoning related, similar route-of-entry or symptoms, etc.) and thus higher

Word2Vec similarity with other drugs compared to the general-concept seedword "drug" (the contexts of "drug" were also often related to consequent injuries resulting from drug abuse in addition to drug (plus alcohol) poisoning).

Using the seedword "diazepam" achieved the highest accuracy, 94%, which could be due to its relatively high DF and small difference within this context (82.6% of narratives associated with "diazepam" were poisoning and allergy related). On the other hand, using the seedword "antibiotic" had a relatively low accuracy (63%), which could be due to its low DF (35) and diverse contexts in the QISU dataset. The injury narratives where "antibiotic" was mentioned were associated with multiple external cause categories (PA_DRUG: 37%; STRUCKCOLLISION: 14%; ANIMAL: 14%; C289: 9%; FALL: 9%; C289EYE: 6%; CUTTING: 6%; HOTOBJ: 3%) and these narratives did not seem to share any obvious pattern. The following are the examples and explanation for the seedwords that had diverse contexts or low DF but higher accuracy than "antibiotic." Although the seedwords"panadol" (11.7%) and "advil" (22%), compared to "antibiotic" (40%), had an even lower percentage of associated documents related to poisoning and allergy, the reason might be because they had either much high DF (584 of total 5,578 drug poisoning injury narratives contained the word"panadol" while only 13 contained "antibiotic") or in obvious pattern in context (more than half of the narratives associated with "advil" had the word "pain"). In addition, the seedword "cocaine" had an even lower DF (7) but slightly higher accuracy (67%) than "antibiotic." This might be because the context of "cocaine" was more consistent – all narratives were 100% related to drug (plus alcohol) poisoning. Thus, an effective seedword for identifying same-hypernym words tend to have higher DF or consistent context. The consistent context can have either small difference within its context (more consistent context among the associated documents or more related content with the concept being identified, such as drug in this case) or a shared, repetitive pattern in context (share a common feature in associated narratives, such as "pain" in more than half of documents associated with "advil").

PA-concept: CHEMICAL

For PA-concept CHEMICAL, the top 100 words most similar to four seedwords "chemical", "detergent", "shampoo", and "paint" are examined respectively. If the word in the top list is not a hyponym of the associated PA-concept, its association with the seedword is identified and coded. Table 7.26 lists the association codes and descriptions, identified from these top word-lists. The association codes include same-hypernym, route, symp, body, other (partial), other (noise), other (descriptive), other ("cause burn"), other ("cause poisoning"), and other ("FB"). The first seven associations are the common artifact of Word2Vec as they also share similar context as the chemical related seedword. The last three associations are also the artifact of the selection of seedword, which will be discussed in more detail later.

Table 7.26: Roles of Most Similar Words to Chemical-related Seedwords

Role	Description
same-hypernym	Share the same hypernym or PA-concept
route	Route of entry. Can be explicit (such as "injected" and "ingestion") or implicit (such as "leaked", "spraying" and "exposure")
symp	Symptom
body	Body part affected
other (partial)	Other: Part of multi-word phrase that shares a hypernym. Often generic and associated with other phrases or categories. For example, "CITRUS" has a high Word2Vec similarity score to "chemical" (as in "CITRUS CLEANER" and "CITRUS AIR FRESHENER"), but it can also related to other plants such as "CITRUS FLOWER", and "CITRUS THORN".
other (noise)	Other: Has no semantic meaning or discriminatory power
other (descriptive)	Other: Describe injury activity such as how, who, when, where, and what involved, including treatment
other ("cause burn")	Injury agent that causes burn
other ("cause poisoning")	Injury agent that causes poisoning
other ("FB")	foreign body that can be ingested or fly / splash into eye; or word related to choking events

PA-concept: CHEMICAL – Seedword "chemical"

First, the top 100 words most similar to the general-concept seedword "chemical" were examined. Table 7.27 list the top 20 most similar words. The word is marked with asterisk if it is the hyponym of the associated PA-concept chemical. If not, I identified the association and assigned an association code and made an additional note if needed (correct spelling of typo, examples of multi-word phrases or injury narratives).

Table 7.27: Top 20 Most Similar Words to the Word *chemical*

Top	CHEMICAL (56%)	DF	Similarity Score	Role
1	acid*	353	0.733757	
2	friction	147	0.721012	other (“cause burn”)
3	splash	398	0.711528	route
4	hydroflouric*	6	0.699146	
5	etc	162	0.698172	other (noise) “FLASH BURN SUNBURN CHEMICAL FRICTION ETC”
6	ester	1	0.695372	other (partial) “CHEMICAL PHOSPHATE ESTER OIL”
7	alkaline*	21	0.68775	
8	caustic*	68	0.685245	
9	sulphuric*	21	0.684523	
10	sunburn	251	0.678871	other (“cause burn”)
11	degreaser*	35	0.675858	
12	lime	39	0.667233	other (partial) “LIME POWDER”
13	scoleded	1	0.663582	symp “SCALDED”
14	ketone*	3	0.651311	
15	hyrdochloric*	1	0.642128	
16	explosion	99	0.640187	route
17	spill	79	0.639069	route
18	splashed	716	0.629447	route
19	flame	95	0.623389	other (“cause burn”)
20	chlorine*	179	0.622697	

The associations identified from the top words to the seedword “chemical” are classified and tabulated in Table 7.28.

Table 7.28: Roles of Top 100 Most Similar Words to *chemical*

Role	%	Words Similar to "chemical"
same-hypernym	56%	*acid, hydroflouric, alkaline, caustic, sulphuric, degreaser, ketone, hyrdochloric, chlorine, liquid, phosphoric, diluted, sulphur, lubricant, soda, preen, nitric, nochlor, bleach, corrosive, hydrochloric, ajax, solution, ammonia, spray, undiluted, deodorant, sanitiser, glowstick, sulphite, hemodent, detergent, petrol, bam, bleech, chemicals, aeroguard, powder, techwash, hydroflpuric, oil, hydrofluoric, solvent, brodon, mould, gas, citroclean, citro, nitrogen, flux, hydroxyflurocarbon, coolant, sulfamic, bifenthrin, polyethylene, perspirant*
other ("cause burn")	16%	*friction, sunburn, flame, steam, ignited, shimmer, flaming, exloded, lighting, exposion, wax, extinquisher, lpg, photographing, weld, flash*
route	10%	*splash, explosion, spill, splashed, sprayed, exposure, spraying, splashing, squirted, plashed*
other (partial)	6%	*based, ester, vapours, product, lime, stain*
other (noise)	6%	*qnd, etc, intol, extinquisher, overfilled, ckecking*
symp	5%	*corrosion, singed, scald, ertherma, scoleded*
other (descriptive)	1%	*smelter*
body	1%	*retinal*

Using the seedword "chemical" to identify the words with the same PA-concept CHEMICAL has the accuracy of 56%, as shown in Table 7.29. Other top words are agents that cause burn (16%), route of entry (10%), part of multi-word phrases (6%), noises (6%), symptoms (5%),descriptive words (1%), and body part (1%). Although most associations are, the only that is not generic artifacts is other ("cause burn"). To explain why the agents that cause burn are considered to be very similar to "chemical" by Word2Vec, I examined the injury narratives where the word "chemical" was mentioned along with their categories for the external injury cause.

Table 7.29: Category Distribution for Documents that Contain the Word *chemical*

Category	Count of Documents that Contain "chemical"	%
ANIMAL	1	0.2%
C289	**158**	**34.5%**
C289EYE	**148**	**31.5%**
CHOKING	3	0.6%
FALL	4	0.8%
FIREFLAME	7	1.5%
HOTOBJ	**42**	**7.1%**
MACHINERY	4	0.8%
MOTORVEHICLE	1	0.2%
PA_CHEMICAL	**101**	**21.2%**
PA_OTHERS	1	0.2%
STRUCKCOLLISION	6	1.3%
Grand Total (DF)	476	

The injury narratives that contain the word "chemical" are often associated with the external cause of chemical burn or exposure (C289; 34.5%), chemical splash in eyes (C289EYE; 31.5%), poisoning due to chemicals (PA_CHEMICAL; 21.2%), and hot object (HOTOBJ; 7.1%). Note that the following string "INJURY BURN FLASH BURN SUNBURN CHEMICAL FRICTION ETC" is often mentioned in the hot object related injury narrative, although the actual injury agent is not associated with chemicals. Three example narratives are listed in the following:

- *"BURN FLASH BURN SUNBURN CHEMICAL FRICTION ETC **HOT WATER** BURN TO BASE OF LEFT THUMB RED AREA SKIN INTACT ADT UTD"*
- *"INJURY BURN FLASH BURN SUNBURN CHEMICAL FRICTION ETC COOKING WITH **HOT FAT** AND TRIPPED PAN BURNS TO INNER THIGHS AND FINGERS COOLING MEASURES AT HOME / WRAPPED IN GLADWRAP TAKEN PAIN RELIEF"*
- *"INJURY BURN FLASH BURN SUNBURN CHEMICAL FRICTION ETC BY TOUCHING **PIZZA** AT LUNCH TIME RUN UNDER WATER AFTER INJURY"*

It can be seen that "hot water", "hot fat" and "pizza" are the cause of the burn injury, instead of "chemical". By reasonably excluding the association between "chemical" and hot object related injury, it can be concluded that the contexts of injury narratives that involve the word "chemical" are mostly related to chemical burn and chemical splash in eye, followed by chemical poisoning. This explains why the words as agents that cause burn have a very high Word2Vec similarity score and account for 16% of top 100 words to "chemical."

PA-concept: CHEMICAL – Seedword "detergent"

A Similar investigation is conducted for the seedword "detergent". Table 7.30 lists the classified words for each associations. The results indicate that the seedword "detergent" is more effective in identifying words with concept of chemical (73% accuracy). In addition to associations that are inevitable artifacts of Word2Vec, 4% of top words are agents that cause poisoning. The result is expected as detergents are also the agent that cause poisoning and share the similar context.

Table 7.30: Roles of Top 100 Most Similar Words to *detergent*

Role	%	Words Similar to "detergent"
same-hypernym	73%	*ajax, bleach, powder, liquid, alkaline, disinfectant, turps, diluted, solution, degreaser, citronella, perfume, batteries, thinner, shampoo, insecticide, thinners, solvent, undiluted, killer, omo, hypochlorite, peroxide, cleanser, deodorant, unleaded, glowstick, mould, dishwashing, soda, chlorine, hydrogen, fragrance, domestos, sulphate, tinsel, preen, chorine, silica, mentholated, alfoil, chemicals, pineoclean, radiant, sodium, bubble, vaporiser, bleech, camphor, lubricant, vicks, deoderant, superglue, reliance, menthol, enamel, flea, acetone, vapour, organophosphate, kerosene, vaporizer, mineral, ratsak, harpic, bam, spray, mortein, glade, rexona, freshner, ratsack, aaa*
other (partial)	19%	*contents, purpose, based, granules, cinnamon, product, contains, clove, dishing, sponge, essential, grime, lavender, flakes, rinse, scented, stain, tanning, weed*
other ("cause poisoning")	4%	*seeds, snail, contraceptive, fiorinal*
route	2%	*sniffed, squirted*
other (descriptive)	1%	*atropine*
other ("FB")	1%	*candy*

The injury narratives that contain the word "detergent" and their categories were examined. Table 7.31 shows the category distribution for these documents. Most documents are associated with the chemical poisoning (56%) and chemical splash (31%). The relatively higher consistency in the context where the "chemical" is mentioned explains the effectiveness of identifying words with chemical concept using the seedword "detergent."

Table 7.31: Category Distribution for Documents that Contain the Word *detergent*

Category	Count of Documents that contain "detergent"	%
ANIMAL	1	1%
C289	3	3%
C289EYE	**30**	**31%**
FALL	1	1%
HOTOBJ	3	3%
PA_CHEMICAL	**54**	**56%**
PA_OTHERS	1	1%
STRUCKCOLLISION	2	2%
Grand Total (DF)	96	

PA-concept: CHEMICAL – Seedword "shampoo"

Furthermore, the effectiveness of using the seedword "shampoo" to identify hyponyms of chemicals was examined. As Table 7.32 shows, the accuracy is 58%, similar to the general-concept seedword "chemical" (56%) but lower than the specific-concept seedword "detergent" (73%). 24% of top words are commonly related words in similar context (i.e., artifacts of Word2Vec) whereas 17% are agents of foreign body with a potential risk of adhering to eyes or breathing problem.

Table 7.32: Roles of Top 100 Most Similar Words to *shampoo*

Role	%	Words Similar to "shampoo"
same-hypernym	58	*ajax, detergent, peroxide, perfume, solution, disinfectant, alfoil, bleach, thinners, insecticide, acetone, superglue, undiluted, lotion, solvent, diluted, thinner, powder, deoderant, toothpaste, mould, bubble, alkaline, turps, vapour, pellets, liquid, tinsel, dettol, degreaser, cleanser, rusk, bubbles, batteries, deodorant, citronella, hairspray, hypochlorite, wipe, vicks, johnsons, killer, bicarb, san, domestos, bleached, hydrogen, bleech, mortein, glo, detol, preen, sanitiser, stingose, chemicals, oils, napisan, flavoured*
other ("FB")	17	*candy, chilli, cordial, crystals, juice, snail, staples, swabs, sweat, sweet, urnie, wrapper, foil, chocking, cooked, matter, pacer, sticky*
other (partial)	14	*contents, biohazard, granules, product, clove, concentrate, dandruff, dissolved, sponge, toner, lavender, containing, rinse, stain*
route	4	*leaked, sniffed, squirted, sucked*
other (descriptive)	4	*rinsed, scooped, spit, stored*
body	2	*lashes, beard*

The word "shampoo" is also a specific-concept seedword and has much lower DF compared to others. As Table 7.33 shows, The injury narratives that contain "chemical" are mostly related to poisoning and allergy (PA_CHEMICAL), or bathroom accidents, including shampoo in eye (C289EYE), or dislocated or hurt shoulder when reaching up for shampoo (C289), spilled shampoo on floor and fell (FALL), or shampoo bottle fell on patient (STRUCKBYCOLLISON). Thus, shampoo can be an agent that causes poisoning and allergy, a foreign body that enters the eye, or an object that can cause people to fall or can fall and hit people. Since the context where the word "chemical" is mentioned is mainly chemical related poisoning and allergy or foreign body in eye, Word2Vec identified the words with similar context: chemicals and agents of foreign body.

Table 7.33: Category Distribution for Documents that Contain the Word *shampoo*

Category	Count of Documents that contain "shampoo"	%
C289	**4**	**9%**
C289EYE	**12**	**28%**
FALL	**8**	**19%**
MACHINERY	1	2%
PA_CHEMICAL	**14**	**33%**
STRUCKCOLLISION	**4**	**9%**
Grand Total (DF)	43	

PA-concept: CHEMICAL – Seedword "paint"

Paint is also a common injury agent that causes poisoning through inhalation and thus can be an interesting example of seedword to examine. As shown in Table 7.34 ,the accuracy of identifying chemicals is 53%, lowest among all chemical-related seedwords but not significantly different from "chemical" or "shampoo". Similar to "shampoo", a great portion of top similar words to "paint" is related to agents of foreign body (19%). In addition to a common-Word2Vec-artifact association (19%), 8% of top words were found to be noisy words, which is the highest among four chemical-related seedwords (3% for "chemical" and 0% for "detergent" and "paint").

Table 7.34: Roles of Top 100 Most Similar Words to *paint*

Role	%	Words Similar to "paint"
same-hypernym	53	*deodorant, spray, enamel, solvent, degreaser, galvit, deodarant, nochlor, thinners, rexona, glowstick, superglue, chemicals, deoderant, ajax, thinner, hairspray, ketone, petrol, epoxy, undiluted, bleech, glue, shampoo, detergent, perfume, glo, peroxide, preen, shave, hyrdochloric, bleach, whiteout, batteries, deisel, alkaline, visine, deodrant, solution, citro, primer, diluted, mould, resin, acid, remover, citronella, hydroflouric, stripper, dulux, disenfectant, chorine, methanol*
other ("FB")	19	*sawdust, woodchip, chipboard, glitter, chemset, dust, routing, particle, particles, flake, matter, foreight, sparks, filings, mixing, fleck, flakes, mix, embrella*
other (noise)	8	*wne, galv, diging, injectors, extinguishers, coupled, intol, stabbe*
other (partial)	7	*based, citrus, contents, geranium, lime, stain, super*
route	6	*spraying, blowing, exploded, squirted, splashed, sprayed*
other (descriptive)	5	*smelter, sewerage, painter, masking, automotive*
body	1	*lashes*
other ("cause burn")	1	*ignited*

The word "paint" is also a specific-concept seedword but has a very high DF, even slightly higher than the word "chemical". Table 7.35 lists the categories associated with the documents that contain "paint" and their corresponding counts and percentages. The word "paint" is associated with a diverse range of categories and contexts, which include: paint as foreign body entering the eye (C289EYE; 20.6%), poisoning due to the inhalation or ingestion of paint or related products such as paint thinner (PA_CHEMICAL; 19%), or any possible panting related injury at work shop (MACHINERY; 12.1%) or other places such as home. Examples of such injuries that occurred at work shop or home include: paint can or scraper

falling on foot or hit foot with them (STRUCKCOLLISION; 14%); sore body, back pain, sprained shoulder or twisted knee when painting (C289; 9.9%); laceration when opening paint tin or using paint scraper (CUTTING; 8.5%); fell from elevation (ladder or roof) or same level on wet paint (FALL; 7.1%).

Table 7.35: Category Distribution for Documents that Contain the Word *paint*

Category	Count of Documents that contain "paint"	%
ANIMAL	3	0.6%
BICYCLE	1	0.2%
C289	**50**	**9.9%**
C289EYE	**104**	**20.6%**
C289FBEAR	2	0.4%
C289FBI	2	0.4%
CHOKING	1	0.2%
CUTTING	**43**	**8.5%**
ELECTRICITY	3	0.6%
FALL	**36**	**7.1%**
FIREFLAME	9	1.8%
HOTOBJ	6	1.2%
MACHINERY	**61**	**12.1%**
MOTORVEHICLE	3	0.6%
PA_ALCOHOL	3	0.6%
PA_CHEMICAL	**96**	**19.0%**
PA_DRUG	9	1.8%
PA_DRUGALCO	3	0.6%
STRUCKCOLLISION	**71**	**14.0%**
Grand Total (DF)	506	

General conclusion for PA-concept CHEMICAL

Four chemical-related seedwords ("chemical", "detergent", "shampoo", "paint") were examined in terms of how effective they are in identifying hyponyms of chemicals. Table 7.36 summarizes the distribution of roles for each chemical-related seedword. Each seedword has its characteristics and particular context. The seedword "chemical" is a generic concept and has the same form as the PA-concept CHEMICAL while the other three are specific chemicals that cause poisoning. Word2Vec measures the distributional similarity between words, thus it quantifies the similarity between words according to how similar their context is. That being said, two words with a high Word2Vec similarity score often share similar context.

Table 7.36: Distribution of Roles and Document Frequency for Top 100 Most Similar Words to Chemical-related Seedwords

	Chemical-related seedword (DF)			
Role	**"chemical"(476)**	**"detergent" (93)**	**"shampoo" (43)**	**"paint" (506)**
same-hypernym (accuracy)	56%	73%	58%	53%
route	10%	2%	4%	6%
symp	5%			
body	1%		2%	1%
other (partial)	6%	19%	14%	7%
other (noise)	3%			8%
other (descriptive)	3%	1%	4%	5%
other ("cause burn")	16%			1%
other ("cause poisoning")		4%		
other ("FB")		1%	17%	19%

In the area of injury, chemicals are the injury agent that can cause poisoning or burn, and enter the eye as foreign body. Such injury narratives often involve the information about route-of-entry, symptom, body part affected, or activity when injury occurred. These "background" words often share the context similar to the word that indicates injury agent, and thus often receive a high Word2Vec similarity. In addition, chemicals can be comprised of multiple words, for example: "cleaning product," "hair spray," "air refresher," "rosemary oil," or "oil based paint." These multi-word phrases often involve a generic word that may be associated with concepts other than chemical, such as "product," "hair," "air," "rosemary," and "based," in previous examples. As a results, such generic words should not be considered as chemicals although they are part of multi-word chemicals and share similar context with them. As an artifact of Word2Vec, several associations were found among the top words similar to chemical-related seedwords, coded as *route*, *symp*, *body*, *other*(*"descriptive"*), and *other*(*"partial"*). In addition to these common Word2Vec-artificat associations, other associations *other*(*"causeburn"*), *other*(*"causepoisoning"*), and *other*(*"FB"*) are the artifacts specifically related to the context and inherent characteristics of selected seedwords.

In addition to being a poisoning-related injury agent, the general-concept seedword "chemical" is also related to chemical burn whereas other specific-concept seedwords "detergent," "shampoo," and "paint" often act as agents of foreign body that enter the eye.

The nature of injuries that involve "shampoo" and "paint" are more complex than "detergent." Shampoo and paint not only imply "poison" but also "physical objects" that are used in the injury-prone place (bathroom) or activity (painting). "Paint" can be put together with other words to create multi-word phrases such as "paint can" and "paint scrapper", which imply different nature of causes: the can of paint is heavy and thus can fall or be struck against, or the paint scrapper is sharp and thus can lacerate human skin. Common injuries in which shampoo and paint often play a role include fall (FALL), overexertion (C289), cutting (CUT-

TING), or struck by (STRUCKCOLLISION). Since "shampoo" and "paint" are associated with diverse injury contexts, they were also found to have more noisy results in Word2Vec. In contrast, "detergent" involves less diverse injury environment and has much more pure nature of context (87% of associated documents are chemical poisoning or entering the eye). Thus, "detergent" was found to be an effective seedword to identify hyponyms of chemicals.

7.2.2 Evaluation of Semantic Grouping Paired with Manual Review in Improving Classification Performance

(i) **Semantic Grouping Method**

By extending the semantic data mining method in Section 7.1.2 (c), the semantic grouping method proposed in this section utilizes the Word2Vec statistical semantics measure to identify agents of poisoning and allergy (PA) related injury (i.e., drug, chemical, alcohol, food, plant) from the injury narratives in the QISU dataset. Specifically, this is accomplished by finding the word that has the highest Word2Vec similarity score with any of the pre-selected seedwords. This method allows for human input in a form of manual review to be incorporated into the machine prediction process. By reviewing the word candidates suggested by Word2Vec, a human reviewer classifies them into different wordlists depending on their PA concept and excludes the ones that do not indicate an injury agent. Given the same-hypernym wordlists for targeted PA concepts, words in the training and test set can be grouped accordingly, either by tagging or mapping in text preprocessing. A machine learning classifier is then trained based on the processed training data and classifies cases from the processed test data.

The procedure of the semantic grouping method is listed below and Figure 7.8:

1. Select the following variables:

   a. Seedwords that are word representatives of PA concepts and can be used to compare words in an operating narrative
   b. Threshold for a word to be considered as the agent word candidate
   c. Stopping criteria (maximum accuracy, number of runs for manual reviews, number of words to review)
2. Train a Word2Vec model on the QISU dataset
3. Prediction by machine: For each word that is not an agent word in an injury narrative:
   a. Calculate the Word2Vec similarity between the word and every seedword
   b. Designate the word whose similarity score is the highest and beyond the threshold as the agent word candidate
4. Verification by human reviewer: Review a unique list of agent word candidates
   a. Classify word candidates into type of agent (drug, chemical, alcohol, food, and plant) or non-agent and create or update the agent and non-agent wordlist
5. Repeat Steps 3 and 4 until the stopping criteria is met
6. Prepare the data set in text preprocessing step using the tagging or mapping strategies.
7. Train a machine learning classifier on the training data and predict the category for test data

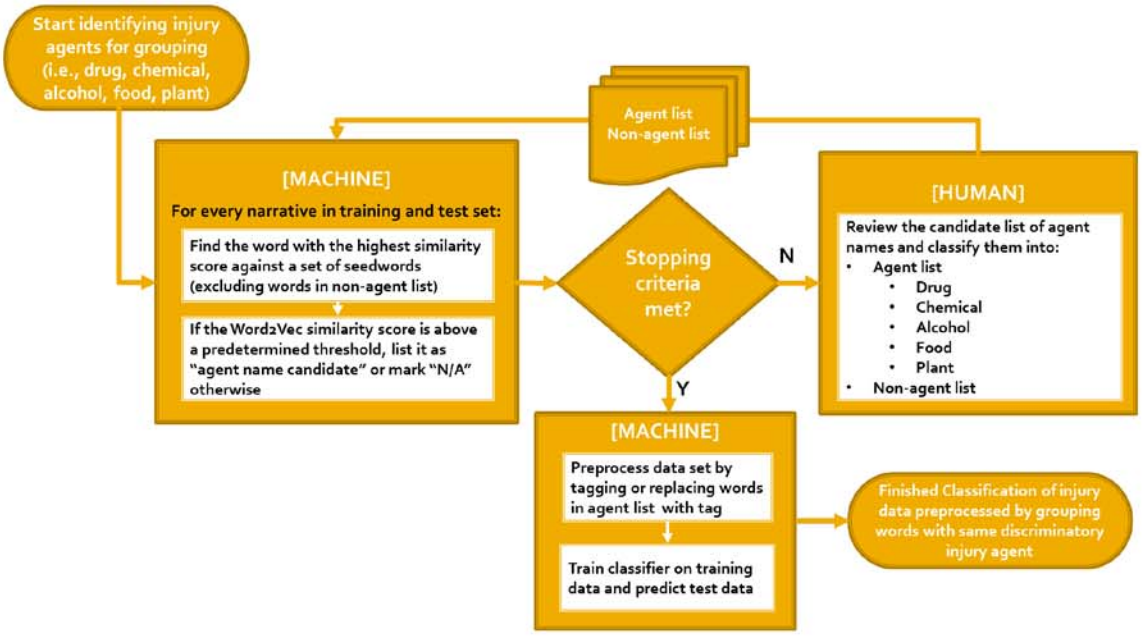

Figure 7.8: Semantic Grouping Method to Identify Agents of Poisoning and Allergy Related Injury

(ii) **Experimental Design**

The proposed semantic grouping method assumes that two words that have a high Word2Vec similarity score tend to share the same semantic concept or hypernym. The current task is to identify words that indicate the PA injury agent by finding the word most similar to a word that carries PA concept (i.e., the concept essential for classifying a PA category).

Based on the result of the exploratory study in Section 7.2.1, seedwords for TA concepts are selected. Table 7.37 shows the seedwords for each PA concept. Each PA concept has one generic-concept seedword, which has the same form of the corresponding PA concept. However, some PA concepts may have one or more specific-concept seedwords because they were found effective in improving the recall of same-hypernym words for the generic-concept seedword in Section 7.2.1.

Table 7.37: PA Concepts and Seedword Lists

PA concept (PA-Hypernym Tag)	**Seedwords**
drug	"drug," "panadol," "ibuprofen," "advil," "cocaine," "antibiotic," "diazepam"
chemical	"chemical,""detergent"
alcohol	"alcohol"
food	"food," "peanut"
plant	"plant"

For this method, every word in an injury narrative is compared against every seedword in Table 7.37 and the Word2Vec similarity score is derived for each possible pair of a word and seedword. For each injury narrative text, the word with the highest Word2Vec similarity score (with any given seedword) that meets the threshold is identified and designated as a word candidate for the corresponding PA injury agent. Those word candidates are called "PA-hyponym" for simplicity and then are grouped, either by tagging or mapping. For tagging, the PA concept (also called "PA-hypernym tag") is added to narratives that contain a PA-hyponym. For mapping, on the other hand, the PA-hyponym is replaced with its corresponding PA-hypernym tag.

The purpose of setting the threshold for the Word2Vec similarity score is to filter out potential cases of false positives because the majority of the QISU dataset is noisy and irrelevant (98% of cases are not PA related). A higher threshold would have a higher precision but lower sensitivity of identifying the potential word candidates, that is, higher accuracy when claimed to be true but lower identification rate of true answers.

In this experiment, three factors, as independent variables, were considered:

- Two grouping strategies
  - Tagging: add a PA-hypernym tag to the narrative that contains a PA-hyponym
  - Tagging: add a PA-hypernym tag to the narrative that contains a PA-hyponym
- Two threshold levels – the lowest Word2Vec similarity score for a word to be considered as the same-hypernym candidate
  - 0.2
  - 0.4
- Three review levels
  - No Review
  - One-time Review
  - Ultimate Review (until no new candidate is suggested for review)

The threshold level and review level together determined the manual review effort. A lower threshold would result in more manual review effort. For the review levels of one-time review and ultimate review, the threshold of 0.2 would always involve more word candidates for review than their 0.4 counterparts. Thus, the review effort can be considered as a combined result of threshold and review levels. By denoting the manual review effort with a concatenated string of threshold and review levels, the order of review effort, from highest to lowest, is: $0.2UltimateReview > 0.4UltimateReview > 0.2OneReview > 0.4OneReview > 0.2NoReview = 0.4NoReview$.

The effect of semantic grouping of PA-concepts and review effort (combined effect of threshold and review levels) was evaluated by the change of classification performance compared to non-grouping. The non-grouping was

standard classification based on the feature space of entire vocabulary without involving any word normalization or grouping method. The performance was measured by the macro-weighted F-measure across all PA categories for each experiment condition based on three classic classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). PA categories are PA_DRUG, PA_ALCOHOL, PA_CHEMICAL, PA_DRUGALCO, PA_FOOD, PA_PLANT, and PA_OTHERS. Unless otherwise noted, the following discussion focuses on the results based on these PA categories as they are the targeted categories for improvement in the current experiment.

I manually identified the words that indicated an agent of injury from PA-related injury narratives in the QISU dataset, and grouped them into different word lists according to their associated injury agent (i.e., drug, chemical, alcohol, food, plant). The classification performance of semantic grouping with different levels of review effort was also compared against the performance of grouping based on the handcrafted dictionary of injury agents.

(iii) **Results and Discussion**

In this experiment, three factors were examined, including: two grouping strategies (tagging and mapping), two thresholds (0.2 and 0.4), and three review levels (no review, one-time review, ultimate review). The threshold and review levels determined the review effort together (0.2NoReview, 0.4NoReview, 0.2OneReview, 0.4OneReview, 0.2UltimateReview, 0.4UltimateReview).

Table 7.38 tabulates the effects of semantic grouping (tagging and mapping) of PA concepts with different levels of review effort (the combination of threshold and review levels) on the macro-averaged F-measure across PA categories for three classifiers and three train-test ratio scenarios. For better visual representations of the results, Figures 7.9-7.14 are shown following Table 7.38.

Table 7.38: Effect of Semantic Grouping for PA Concepts

Train-Test Ratios	Review Effort (Threshold + Review Levels)	MNB		SVM		LR		Overall
		Tagging	Mapping	Tagging	Mapping	Tagging	Mapping	
1:9	0.2 No Review	-5.8%	-12.7%	-3.1%	-13.9%	-2.5%	-12.7%	-8.4%
	0.2 One-time Review	7.3%	0.8%	8.5%	7.1%	11.4%	10.4%	7.6%
	0.2 Ultimate Review	7.5%	0.9%	9.4%	8.3%	12.3%	11.5%	8.3%
	0.4 No Review	1.2%	-6.2%	1.1%	-4.8%	2.8%	-2.7%	-1.5%
	0.4 One-time Review	7.8%	1.6%	8.0%	7.0%	10.6%	10.0%	7.5%
	0.4 Ultimate Review	8.0%	1.8%	8.5%	7.7%	11.2%	10.7%	8.0%
	Handcrafted List	*7.3%*	*0.9%*	*9.4%*	*8.0%*	*12.5%*	*11.2%*	*8.2%*
1:1	0.2 No Review	-4.9%	-13.8%	-3.7%	-18.1%	-3.3%	-18.5%	-10.4%
	0.2 One-time Review	5.4%	-0.6%	4.1%	2.2%	6.4%	4.9%	3.8%
	0.2 Ultimate Review	5.6%	-0.5%	4.8%	2.8%	7.2%	5.7%	4.3%
	0.4 No Review	2.1%	-6.7%	-0.9%	-8.3%	-0.2%	-7.0%	-3.5%
	0.4 One-time Review	5.5%	0.0%	3.8%	2.0%	5.7%	4.3%	3.6%
	0.4 Ultimate Review	5.6%	0.1%	4.1%	2.5%	6.2%	4.9%	3.9%
	Handcrafted List	*5.7%*	*-0.5%*	*5.4%*	*2.6%*	*7.7%*	*5.8%*	*4.4%*
9:1	0.2 No Review	-4.2%	-14.0%	-2.9%	-18.7%	-2.2%	-19.1%	-10.2%
	0.2 One-time Review	5.5%	-0.2%	4.2%	1.7%	6.1%	4.1%	3.6%
	0.2 Ultimate Review	5.6%	0.1%	4.6%	2.2%	7.0%	5.3%	4.1%
	0.4 No Review	1.1%	-6.0%	-0.3%	-8.9%	-0.1%	-8.0%	-3.7%
	0.4 One-time Review	5.5%	0.0%	3.8%	1.9%	5.4%	3.7%	3.4%
	0.4 Ultimate Review	5.6%	0.5%	4.1%	2.1%	6.1%	4.1%	3.8%
	Handcrafted List	*5.6%*	*0.1%*	*4.6%*	*1.4%*	*7.4%*	*4.9%*	*4.0%*

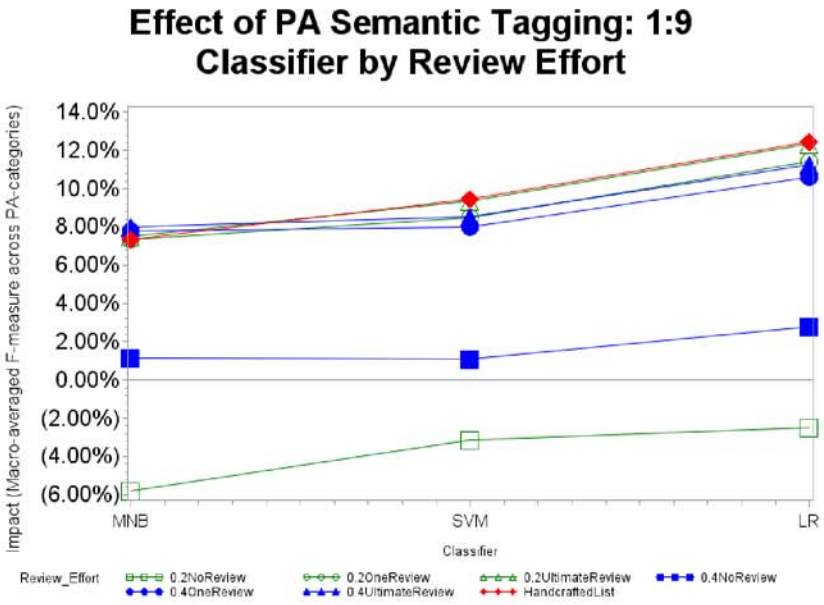

Figure 7.9: Effect of PA Semantic Tagging and Review Effort at Train-Test Ratio of 1:9

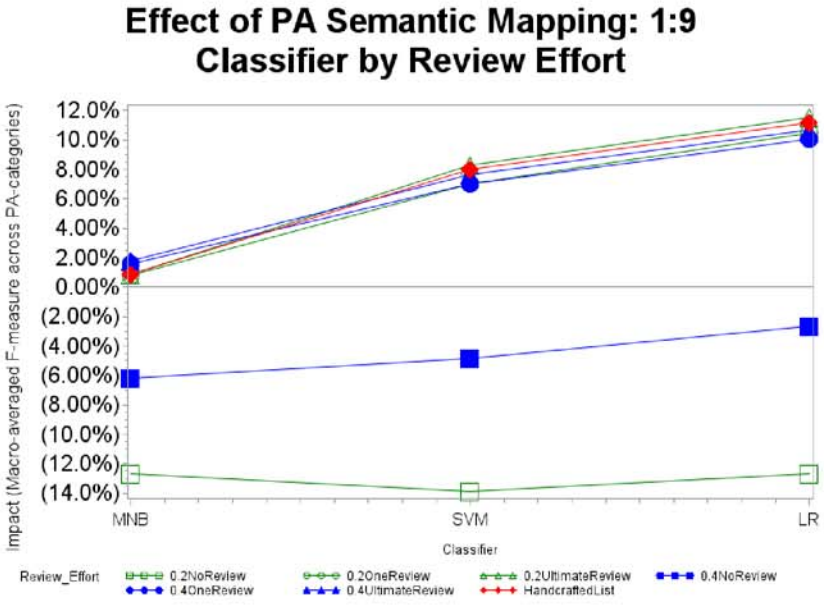

Figure 7.10: Effect of PA Semantic Mapping and Review Effort at Train-Test Ratio of 1:9

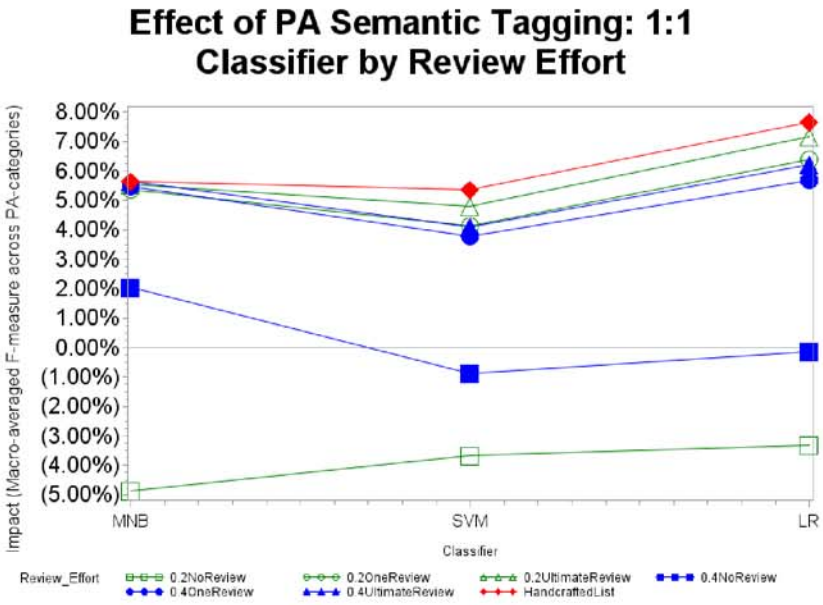

Figure 7.11: Effect of PA Semantic Tagging and Review Effort at Train-Test Ratio of 1:1

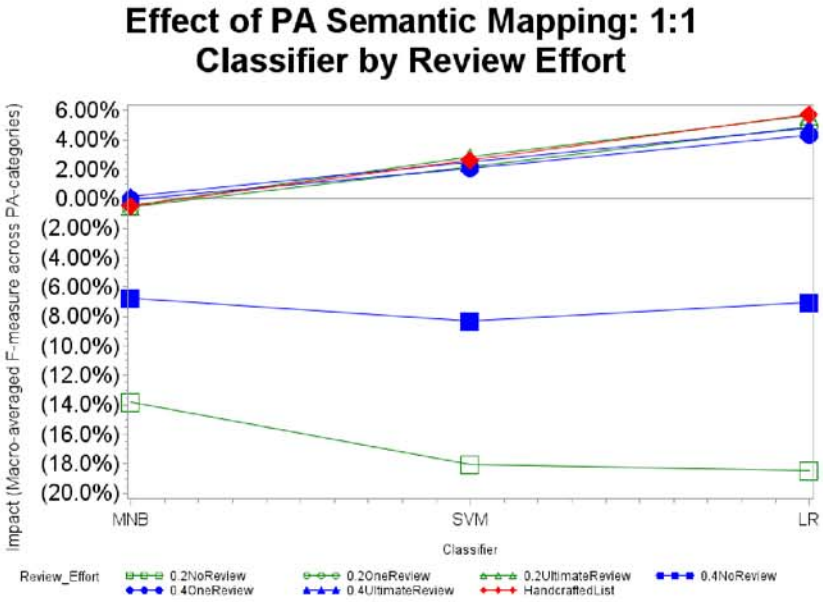

Figure 7.12: Effect of PA Semantic Mapping and Review Effort at Train-Test Ratio of 1:1

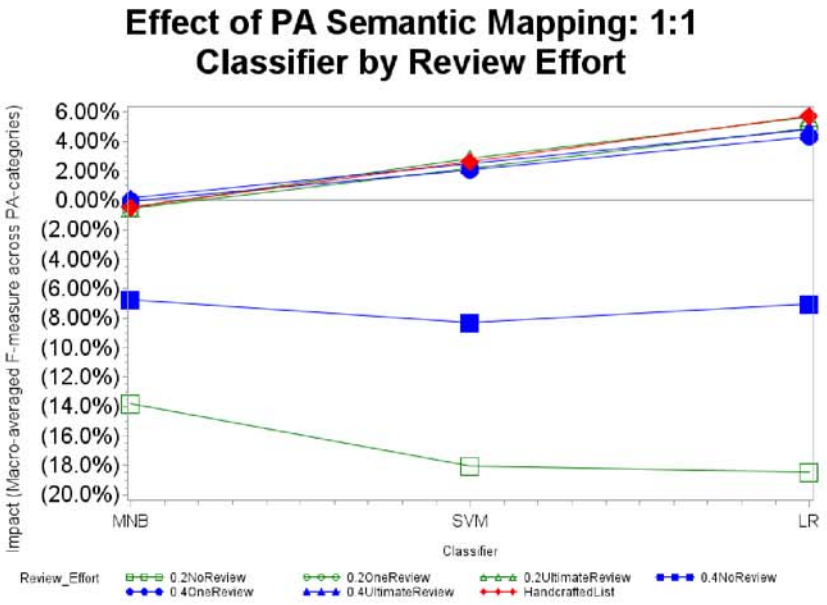

Figure 7.13: Effect of PA Semantic Tagging and Review Effort at Train-Test Ratio of 9:1

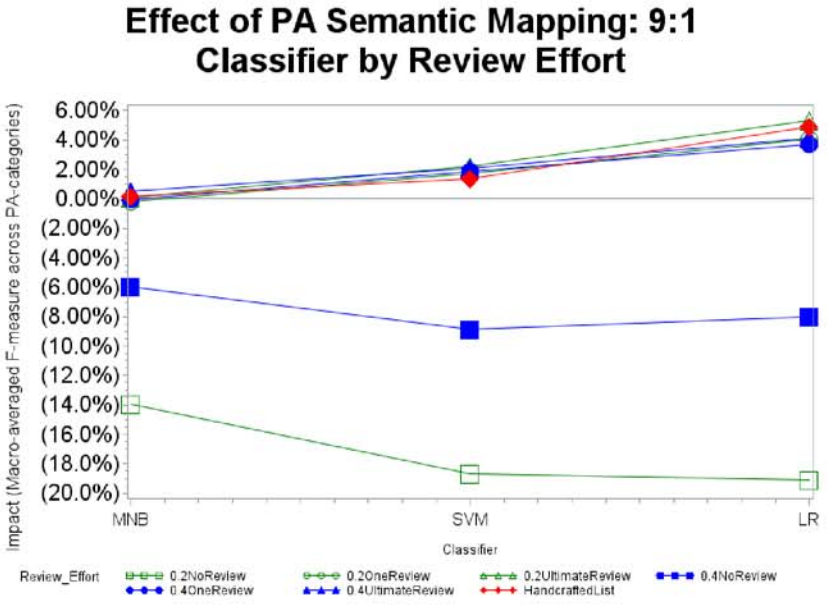

Figure 7.14: Effect of PA Semantic Mapping and Review Effort at Train-Test Ratio of 9:1

In Figures 7.9-7.14, three line clusters can be observed, which are, from top to bottom: (1) grouping with review (the lines that represent one-time review and ultimate review tend to overlap or gather together), (2) grouping without review based on a higher threshold 0.4NoReview, and (3) grouping without review based on a lower threshold 0.2NoReview. This suggests three distinct patterns of semantic grouping effect. There following paragraphs discuss the effect from the perspectives of grouping with and without review.

Overall, the semantic grouping, both mapping and tagging, with review shows positive effects on the overall classification performance of PA categories in terms of the macro-averaged F-measures for all three classifiers in all three train-test scenarios. Figures 7.9-7.14 show that the lines represented for grouping with review (0.2OneReview, 0.4OneReview, 0.2UltimateReview, 0.4UltimateReview) are all above the zero reference line. In addition, semantic grouping was found to have a similar impact as the grouping using a handcrafted dictionary.

When the semantic grouping was performed without review, the classification performance was negatively influenced except for the following conditions. With the tagging strategy, no review with a higher threshold (0.4NoReview) improved the macro-averaged F-measure of PA categories by around 1% for MNB and SVM and 3% for LR when the train-test ratio is 1:9 and by 1% to 2% for MNB in all three train-test ratio scenarios. Other than those exceptions, conducting the semantic grouping without review showed a negative impact. Semantic tagging without review and with a lower threshold (0.2NoReview) decreased the macro-averaged F-measure by 3% to 6%. As for using mapping as the grouping strategy, the review effort 0.4NoReview decreased the macro-averaged F-measure by 3% to 8% and 0.2NoReview decreased it by 13% to 19%.

Semantic grouping can have different effects depending on the selection of thresholds and grouping strategies. The effect of threshold is summarized in Table 7.39, which is the impact difference between the two threshold levels for different experimental conditions. Since the impact was calculated by subtracting the impact of threshold 0.4 from the impact of threshold 0.2, positive figures indicate the positive impact of using a lower threshold (0.2) and negative figures indicate the negative impact of using a lower threshold (0.2). With review (OneReview and UltimateReview), using a lower threshold was found to have a slightly better impact for SVM and LR (range: 0.1% to 1.2%; mean: 0.58%) but a slightly worse impact for MNB (range: -0.9% to 0%; mean: -0.38%) than its higher threshold counterpart.

The impact difference between two threshold levels was more significant for the condition of no review. Without review, using a lower threshold (0.2) was found to have a more severe negative impact than its higher threshold counterpart (0.4) (range: -2.6% to -11.5%; mean: -6.8%). Figures 7.9-7.14 consistently show that the line of 0.2NoReview with hollow squares is always below the line of 0.4NoReview with filled squares. The overall negative impact of semantic grouping on the macro-averaged F-measure across PA categories for 0.2NoReview and 0.4NoReivew was: -8.4% and -1.5% when the train-test ratio is 1:9, -10.4% and -3.5% when the ratio is 1:1, and -10.2% and-3.7% when the ratio is 9:1 (in the Overall column of Table 7.38). In short, the grouping without manual review with a lower threshold (0.2NoReview) decreased the macro-averaged F-measure from a range of 2.6% to 11.5% (mean: 6.8%), increasing the negative impact by a factor of 3 to 5 compared to its higher threshold counterpart (0.4NoReview).

Table 7.39: Effects of Using a Lower Threshold (0.2) Compared to Higher Threshold (0.4) on Impact of Semantic Grouping and Review Level

Train-Test Ratios	Review Level	MNB		SVM		LR		Overall
		Tagging	Mapping	Tagging	Mapping	Tagging	Mapping	
1:9	No Review	-7.00%	-6.50%	-4.20%	-9.10%	-5.30%	-10.00%	-6.90%
	One-time Review	-0.50%	-0.80%	0.50%	0.10%	0.80%	0.40%	0.10%
	Ultimate Review	-0.50%	-0.90%	0.90%	0.60%	1.10%	0.80%	0.30%
1:1	No Review	-7.00%	-7.10%	-2.80%	-9.80%	-3.10%	-11.50%	-6.90%
	One-time Review	-0.10%	-0.60%	0.30%	0.20%	0.70%	0.60%	0.20%
	Ultimate Review	0.00%	-0.60%	0.70%	0.30%	1.00%	0.80%	0.40%
9:1	No Review	-5.30%	-8.00%	-2.60%	-9.80%	-2.10%	-11.10%	-6.50%
	One-time Review	0.00%	-0.20%	0.40%	-0.20%	0.70%	0.40%	0.20%
	Ultimate Review	0.00%	-0.40%	0.50%	0.10%	0.90%	1.20%	0.30%

In addition to the threshold level, different grouping strategies (mapping or tagging) also influenced the level of negative impact under the condition of no review effort. Semantic mapping without review always caused more severe negative impact than semantic tagging without review. In Figures 7.9-7.14, the lines of 0.2NoReview with hollow squares and 0.4NoReview with solid squares are always much lower in Figures 7.10, 7.12, and 7.14 (effect of semantic mapping) than in Figures 7.9, 7.11, and 7.13 (effect of semantic tagging). Table 7.40 tabulates the numerical impact difference between tagging and mapping (tagging – mapping). All figures are positive, implying that semantic tagging is more effective than semantic mapping (range: 2.5% to 14.17%; mean: 4.96%) as a grouping strategy for improving the classification performance of PA categories.

Table 7.40: Effects of Semantic Tagging Compared to Semantic Mapping

Train-Test Ratios	Review Effort (Threshold + Review levels)	Impact (Tagging – Mapping)			
		MNB	SVM	LR	Overall
1:9	0.2 No Review	6.90%	10.80%	10.20%	9.30%
	0.2 One-time Review	6.50%	1.40%	1.00%	2.97%
	0.2 Ultimate Review	6.60%	1.10%	0.80%	2.83%
	0.4 No Review	7.40%	5.90%	5.50%	6.27%
	0.4 One-time Review	6.20%	1.00%	0.60%	2.60%
	0.4 Ultimate Review	6.20%	0.80%	0.50%	2.50%
	Handcrafted List	6.40%	1.40%	1.30%	3.03%
1:1	0.2 No Review	8.90%	14.40%	15.20%	12.83%
	0.2 One-time Review	6.00%	1.90%	1.50%	3.13%
	0.2 Ultimate Review	6.10%	2.00%	1.50%	3.20%
	0.4 No Review	8.80%	7.40%	6.80%	7.67%
	0.4 One-time Review	5.50%	1.80%	1.40%	2.90%
	0.4 Ultimate Review	5.50%	1.60%	1.30%	2.80%
	Handcrafted List	6.20%	2.80%	1.90%	3.63%
9:1	0.2 No Review	9.80%	15.80%	16.90%	14.17%
	0.2 One-time Review	5.70%	2.50%	2.00%	3.40%
	0.2 Ultimate Review	5.50%	2.40%	1.70%	3.20%
	0.4 No Review	7.10%	8.60%	7.90%	7.87%
	0.4 One-time Review	5.50%	1.90%	1.70%	3.03%
	0.4 Ultimate Review	5.10%	2.00%	2.00%	3.03%
	Handcrafted List	5.50%	3.20%	2.50%	3.73%
	Average Impact Difference Between Tagging and Mapping	6.54%	4.32%	4.01%	4.96%

The following discussion focuses on the combined effects of grouping (tagging and mapping) and review (with and without review). The four experiment conditions are denoted with "TaggingWithReview," "TaggingWORe-

view," "MappingWithReview," and "MappingWOReview." Figures 7.15-7.17 display the combined effects of grouping and review on the macro-averaged F1 measure across PA categories for three classifiers (MNB, SVM, LR), with one graph for each train-test scenario: 1:9, 1:1, and 9:1.

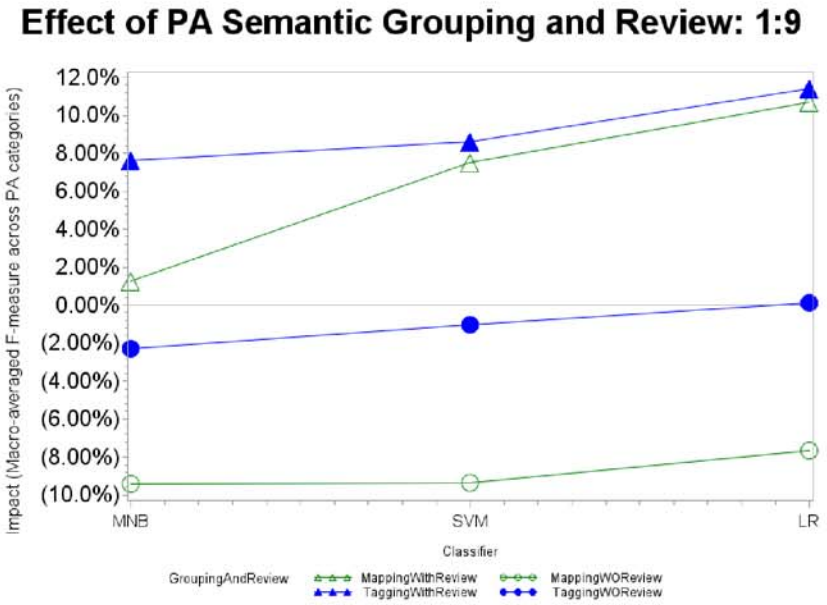

Figure 7.15: Combined Effect of Semantic Grouping and Review at Train-Test Ratio of 1:9

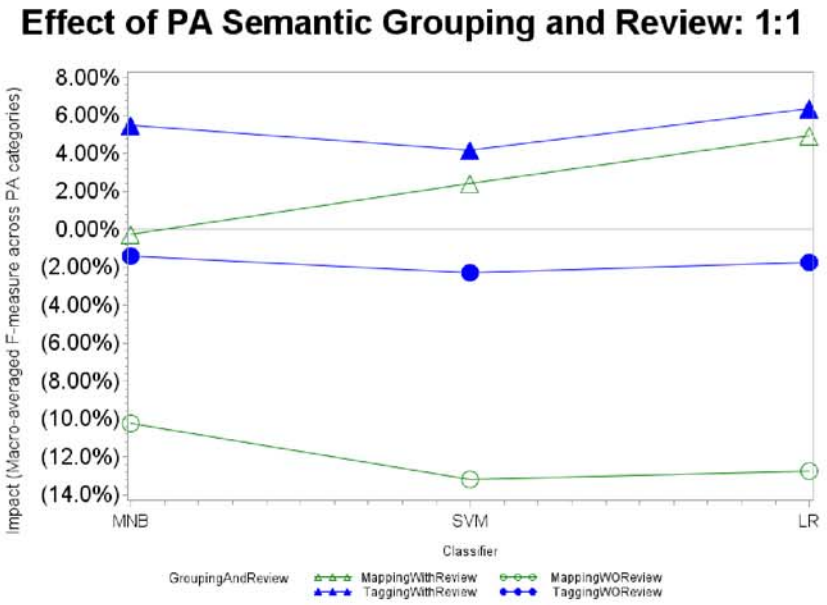

Figure 7.16: Combined Effect of Semantic Grouping and Review at Train-Test Ratio of 1:1

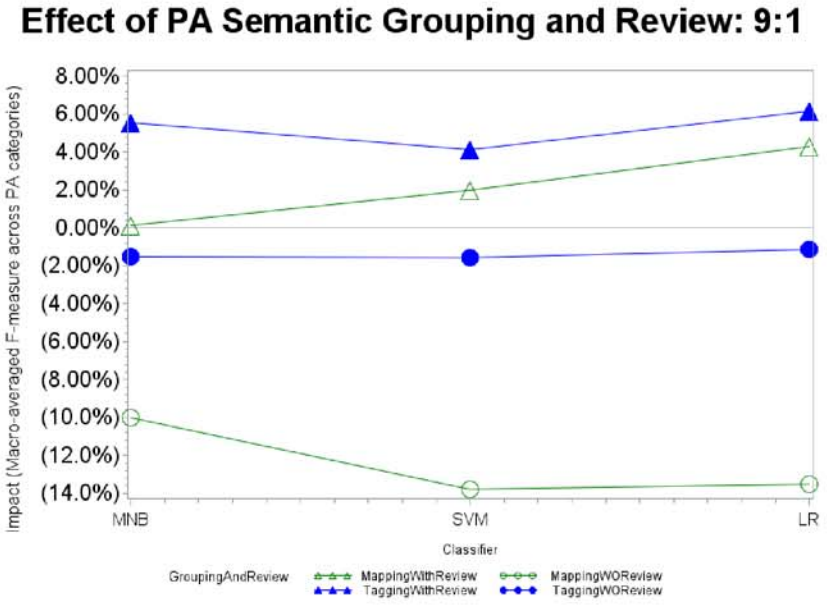

Figure 7.17: Combined Effect of Semantic Grouping and Review at Train-Test Ratio of 9:1

Figures 7.15-7.17 consistently demonstrate the same order of lines from top to bottom, which is: TaggingWithReview, MappingWithReview, TaggingWOReview, and MappongWOReview. The lines of TaggingWithReview and MappingWithReview are always above the zero reference line while the lines of TaggingWOReview and MappongWOReview are always below it. The results suggest that, with a low to moderate threshold (0.2 or 0.4) that filters out some potential cases of false positives, semantic grouping with review can improve the classification performance (range: 3.71% to 7.85%; mean: 5.14%) whereas the grouping without review tends to decrease the performance (range: -4.94% to -6.94%; mean: -6.28%). Table 7.41 tabulates the corresponding numerical effects of grouping and review for three classifiers in three train-test ratio scenarios.

In addition, the lines with solid triangles and circles (tagging with/without review) are always higher than lines with hollow triangles and circles (mapping with/without review). This, again, implies the semantic tagging is more effective than mapping in improving the classification performance of PA categories. That being said, adding extra informative tags to original injury narratives can have a more positive effect on classification performance than replacing words with generic-concept tags in injury surveillance data.

Table 7.41: Effect of Semantic Grouping: Grouping (Tagging and Mapping) and Review (With and Without Review)

Train-Test Ratios	With/Without Review	MNB		SVM		LR		Overall
		Tagging	Mapping	Tagging	Mapping	Tagging	Mapping	
1:9	WithReview	-2.30%	-9.45%	-1.00%	-9.35%	0.15%	-7.70%	-4.94%
	WOReview	7.65%	1.28%	8.60%	7.53%	11.38%	10.65%	7.85%
1:1	WithReview	-1.40%	-10.25%	-2.30%	-13.20%	-1.75%	-12.75%	-6.94%
	WOReview	5.53%	-0.25%	4.20%	2.38%	6.38%	4.95%	3.86%
9:1	WithReview	-1.55%	-10.00%	-1.60%	-13.80%	-1.15%	-13.55%	-6.94%
	WOReview	5.55%	0.10%	4.18%	1.98%	6.15%	4.30%	3.71%

Next, the effect of manual review on the impact of PA semantic grouping was examined, that is, how much semantic grouping impacted the classification performance as the manual review effort increased. Figure 7.18 demonstrates the relationship between the review level (NoReview, OneReview, UltimateReview) and the combined effect of grouping strategies and thresholds (Mapping0.2, Mapping0.4, Tagging0.2, Tagging0.4) in terms of their impact on the macro-averaged F-measure across PA categories for three classic classifiers and three train-test ratio scenarios.

The classification improvement due to the one-time review was examined from the perspectives of threshold and grouping strategy. For the threshold, one-time review tended to improve a lower threshold level (0.2) more than a higher threshold (0.4) due to the fact that more word candidates were reviewed. Figure 7.18 visually shows that the impact increase from NoReview to OneReview was always greater for the threshold of 0.2 (the lines with solid and hollow triangles) than 0.4 (the lines with solid and hollow circles).

For the grouping strategy, on the other hand, one-time review tends to improve the semantic mapping (the lines with hollow triangles and circles)

more than the tagging (the lines with solid triangles and circles) although the final impact of mapping is not as high as the tagging.

In addition, the impact difference between tagging and mapping after one-time review is much larger for MNB (6%) than for SVM and LR (2%). For MNB, semantic mapping even with manual review is ineffective in improving the classification performance because the slight positive improvement (2%) was found only when the availability of training set is limited (at train-test ratio of 1:9).

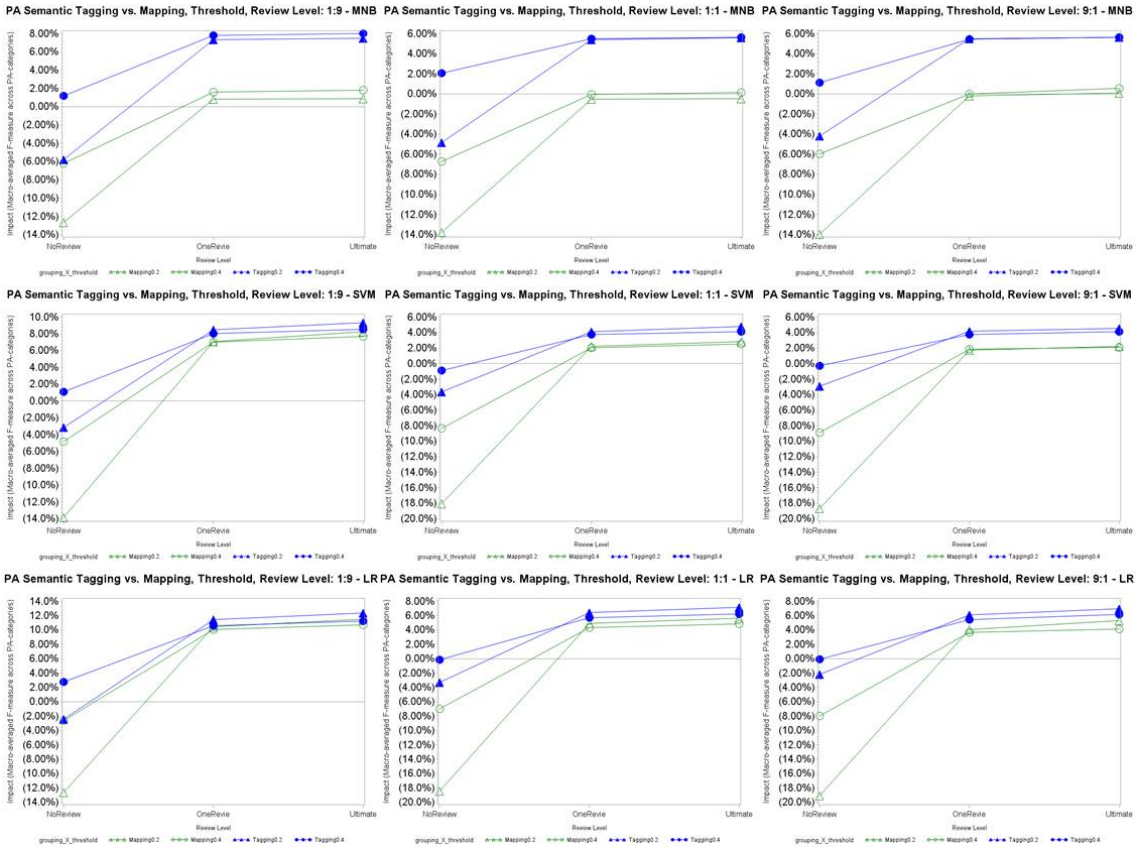

Figure 7.18: Effect of Manual Review on Impact of PA Semantic Grouping

In the injury and safety research, the route-of-entry information also provides insight into the nature of injury. Examples of routes of entry include ingestion, injection, inhalation, and intake. As the routes-of-entry also have

many same-hyponym morphological variants, I tested if the semantic grouping of routes-of-entry would further improve the classification performance of the grouping of injury agents.

The previous experiments showed that semantic tagging for injury agents with one-time review and moderate threshold (0.4) had the best benefit-cost (classification improvement-manual review effort) ratio among all experiment conditions. Thus, this experiment condition was selected as the control group to compare against the treatment groups (semantic grouping for route-of-entry).

The wordlist of routes of entry was developed by manually examining the poisoning and allergy related injury narratives. Two semantic grouping strategies were tested. For mapping, words in the route-of-entry wordlist were replaced with the word "route." For tagging, on the other hand, the word "route" was added to the narratives that contained any word in the route-of-entry wordlist.

The effect of semantic tagging and mapping for route-of-entry was tested based on the classification performance of three classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). The initial results showed that the effect of semantic mapping and grouping of routes-of-entry was not significant. The possible reason could be that grouping all routes of entry into one united concept decreased the precision and words lost their specificity. Thus, the routes of entry were further grouped into sub-categories based on their nature. Table 7.43 lists the words for each route of entry class. Instead of grouping into "route", these words were grouped into specific route-of-entry tags.

Table 7.42: Route-of-Entry Wordlist for Semantic Grouping

Route of Entry	Words to Be Grouped
ingest	*chewed, chewing, swallow, swallowing, swollowed, swollowing, swallowed, swallowed, swollwed, swalllowed, swolled, ingest, ingested, ingesting, ingestion, injested, injestion*
inhale	*inhale, inhalation, inhaled, inhaling, smell, smelled, smelling, smells, sucked, sucking, sniff, sniffed, sniffing*
consume	*intake, intpoxicated, intoxicated, intoxication, consumed, consuming, consumption, consummed*
eat	*ate, eat, eaten, eats, eating*
drink	*drank, drink, drinking, drinks, drinked*
splash	*splash, splashed, splashing, sprayed, spraying, pouring, dripped, poured, spilt, spill, spilled, dripped, exposure, squirted*
swig	*swig, swigs, sipping, licking, licked*

Figures 7.19-7.21 graphically demonstrate the effect of the route-of-entry grouping on classification performance based on the semantic tagging for injury agents with one-time review and threshold of 0.4. In these figures, the original experimental condition of semantic tagging of injury agents is denoted with the treatment label "control" (lines with solid diamonds), and the additional semantic grouping for route-of-entry are denoted with *"routeMapping"* (lines with solid circles) and *"routeTagging"* (lines with solid triangles) for two grouping strategies.

The results suggested that the classification performance of semantic tagging for injury agents can be further improved by the additional semantic tagging of routes-of-entry *("routeTagging")* for MNB and LR, though not for SVM in all train-test ratio scenarios. However, additional semantic mapping for route-of-entry *("routeMapping")* decreased the performance of semantic tagging for injury agents.

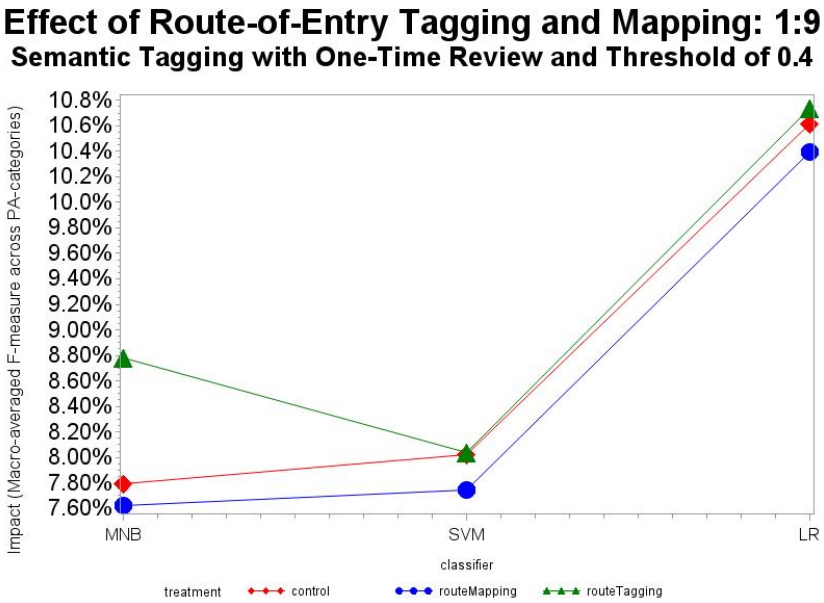

Figure 7.19: Effect of Route-of-Entry Tagging and Mapping at Train-Test Ratio of 1:9

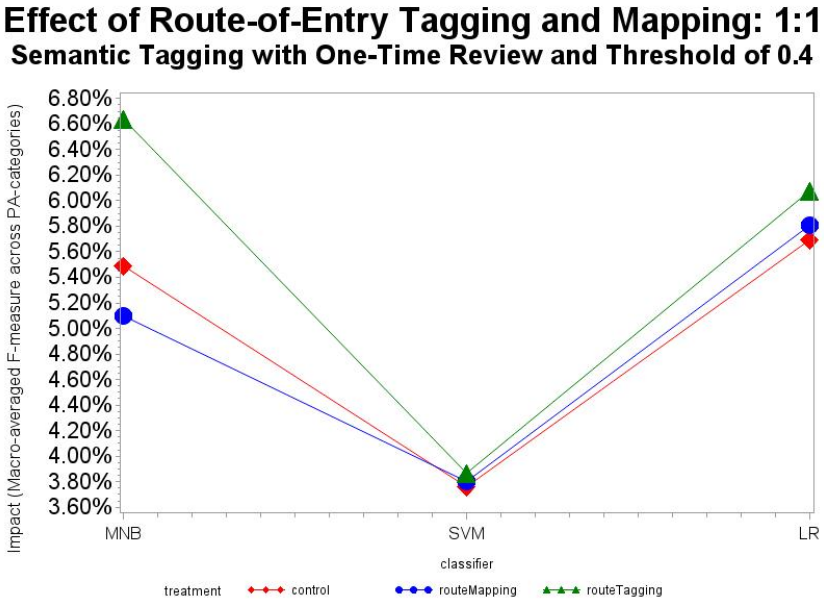

Figure 7.20: Effect of Route-of-Entry Tagging and Mapping at Train-Test Ratio of 1:1

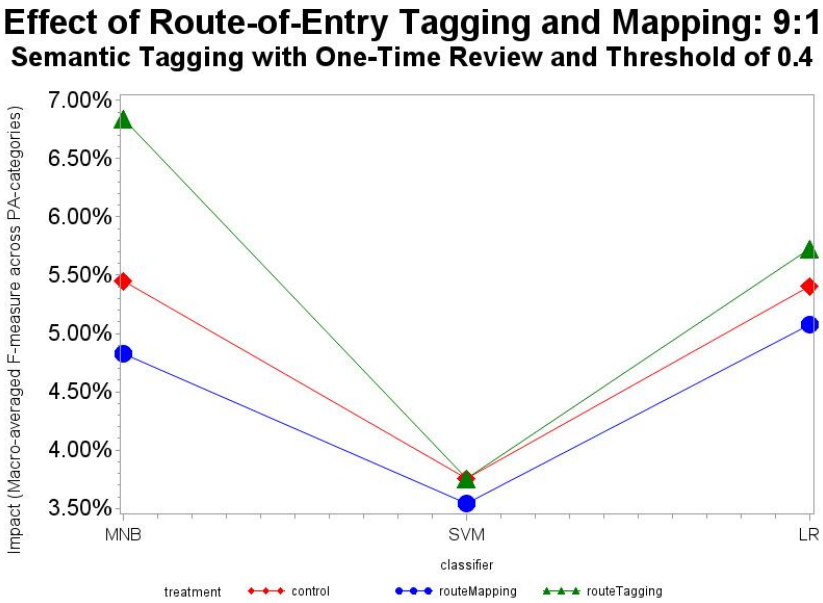

Figure 7.21: Effect of Route-of-Entry Tagging and Mapping at Train-Test Ratio of 9:1

(iv) **Summary**

In Section 7.2.2, I tested if the semantic grouping of poisoning and allergy concept (PA-concept) can improve the classification performance of the associated PA categories. The targeted PA-concepts include drug, chemical, alcohol, food, and plant. As the parameters of the semantic grouping method, two grouping strategies (tagging and mapping), two threshold levels ("moderate": 0.4 and "low": 0.2) and three review levels (no review, one-time review, and ultimate review) were also examined. The effectiveness of semantic grouping of PA-concept was evaluated based on the classification performance of MNB, LR, and SVM in three train-test ratio scenarios (1:9, 1:1, 9:1).

Since this study used a low-to-moderate threshold, manual review effort seems to be required for correcting a high rate of false positive. When the semantic grouping of PA-concept was done with manual review, a positive effect on the classification performance of PA categories was observed. Using

different grouping strategies resulted in different effects. Semantic tagging with review (range: 3.8% to 12.3%; mean: 6.6%) was more effectiveness than mapping with review (range: -0.6% to 11.5%; mean: 3.7%). However, when the semantic grouping was performed without manual review, the classification performance was decreased and mapping (range: -2.7% to -19.1%; mean: -11.1%) had a more severe negative impact than tagging (range: 2.8% to -5.8%; mean: -1.4%). Only when the training data was limited or when MNB was used for classification, the semantic tagging (not for mapping) without review had a slightly positive effect (roughly 1%).

One-time review (0.2: review 50% of the vocabulary; 0.4: 13%) was considered more practical or realistic, as ultimate review (0.2: review 70% of the vocabulary; 0.4: 19%) requires much more effort to achieve a marginal improvement (range: 0.1% to 1.2%; mean: 0.5%). Ultimate review was able to achieve a similar level of positive impact on classification as the handcrafted dictionary. Semantic Tagging with one-time review and a moderate threshold (0.4) was found to have the best cost-benefit ratio (review effort – improved classification performance) among all the experimental conditions due to the doubled review effort and insignificant improvement using a low threshold (0.2). Furthermore, semantic tagging for route-of-entry was able to further slightly improve the classification performance of PA categories for MNB (1%) and LR (0.2%), though not SVM, in addition to semantic tagging of PA-concept.

7.3 Semantic Tagging with Word2Vec in Improving Classification Performance

The proposed semantic grouping method requires a seedword as an input argument to identify words with similar concept. The selection of seedwords is important because the concept they indicate should have discriminatory power for classification. Previous experiments in Section 7.2 demonstrated that grouping of

words with similar and discriminatory concepts (injury agents and routes of entry) can improve the classification performance of poisoning and allergy (PA) related categories. In Section 7.3, I extended the grouping method to other categories, not limited to PA categories. Due to the large volume of predictive categories, manually selecting seedwords for all categories can be time-consuming. From the literature, feature selection methods have been commonly applied to identify discriminatory features for classification. Thus, I examined the effectiveness of classic feature selection methods in identifying seedwords for semantic grouping in order to improve the classification performance.

In this experiment, two factors, as independent variables, were considered:

- Six feature selection methods
  - Mutual information (MI)
  - Odds Ratio (OR)
  - Chi-square statistics (CHI)
  - Coefficient matrix of MNB classifier (MNB)
  - Coefficient matrix of SVM classifier (SVM)
  - Coefficient matrix of LR classifier (LR)
- Five N possibilities for top N highest weight words for each category
  - 10
  - 20
  - 30
  - 40
  - 50

As an illustrative example, some of the top seedwords identified by CHI for the PA_ALCOHOL category included “bourbon,” “turps” (i.e., Australian English

meaning alcoholic drinks), "shots".Table 7.43 lists the words that were statistically similar to these three seedwords, along with their DF, Word2Vec score, and role if not a chemical. Some of these identified words, though not alcohols, were related to alcohols in some ways (e.g. measurement unit for alcohols and route-of-entry or injury agent for poisoning).

Table 7.43: Example of Words Being Grouped by Type-S Semantic Tagging

word	DF	W2V_score	Type-S Tag	Role
vodka	133	0.8959	bourbon	
cask	20	0.8945	bourbon	Unit
scotch	34	0.8762	bourbon	
rum	181	0.8631	bourbon	
whisky	11	0.8619	bourbon	
cans	75	0.8504	bourbon	Unit
cones	42	0.8469	shots	Drug
casks	21	0.8404	shots	Unit
bleach	180	0.8272	turps	Chemical
smoked	38	0.8241	shots	Route of entry
amounts	56	0.8159	bourbon	Unit
whiskey	9	0.8125	bourbon	
rums	31	0.812	shots	
moselle	25	0.8096	shots	
udl	8	0.8093	shots	
dishwashing	57	0.8022	turps	Chemical
chorine	2	0.8013	turps	Chemical

Similar to previous experiments, the effect of semantic grouping was evaluated based on the classification performance difference between grouping and non-grouping (i.e., the standard classification based on the feature space of the entire vocabulary without any grouping). The classification performance was measured

by the macro-averaged F-measure across all 30 categories classified by MNB, SVM, and LR in three train-test ratio scenarios (1:9, 1:9, 9:1).

Table 7.44 lists the numerical effect of semantic grouping, based on the Top N seedwords identified by six feature selection methods for each category, on the classification performance using three classic classifiers in three train-test ratio scenarios.

Table 7.44: Effect of Semantic Grouping Pairing with Six Feature Selection Methods for Selecting Top N Discriminatory Seedwords Per category

Feature Selection	Top N	1:9			1:1			9:1			Overall
		MNB	SVM	LR	MNB	SVM	LR	MNB	SVM	LR	
CHI	10	1.72%	0.64%	0.70%	1.40%	0.24%	0.37%	1.02%	0.23%	0.31%	0.76%
	20	1.70%	0.67%	0.61%	1.36%	0.34%	0.22%	1.09%	0.24%	0.30%	0.75%
	30	1.66%	0.68%	0.63%	1.62%	0.28%	0.23%	1.14%	0.16%	0.29%	0.76%
	40	1.65%	0.69%	0.60%	1.57%	0.19%	0.21%	1.15%	0.13%	0.33%	0.75%
	50	1.60%	0.64%	0.54%	1.43%	0.24%	0.29%	1.18%	0.08%	0.29%	0.72%
LR	10	1.61%	0.68%	0.68%	1.10%	0.06%	0.18%	0.91%	0.19%	0.22%	0.69%
	20	1.60%	0.66%	0.67%	0.79%	0.16%	0.31%	0.84%	0.18%	0.25%	0.68%
	30	1.64%	0.65%	0.64%	0.89%	0.20%	0.22%	0.74%	0.16%	0.15%	0.64%
	40	1.66%	0.66%	0.65%	0.92%	0.30%	0.28%	0.78%	0.19%	0.06%	0.65%
	50	1.64%	0.67%	0.64%	0.92%	0.28%	0.25%	0.73%	0.20%	0.10%	0.65%
MI	10	1.55%	0.64%	0.68%	0.82%	0.27%	0.18%	0.82%	0.05%	0.09%	0.62%
	20	1.61%	0.67%	0.63%	0.82%	0.24%	0.28%	0.71%	0.13%	-0.04%	0.60%
	30	1.55%	0.64%	0.61%	0.83%	0.23%	0.30%	0.72%	0.18%	-0.06%	0.59%
	40	1.52%	0.64%	0.63%	0.81%	0.32%	0.28%	0.70%	0.11%	-0.10%	0.57%
	50	1.52%	0.67%	0.58%	0.83%	0.29%	0.29%	0.70%	0.12%	-0.13%	0.57%
MNB	10	1.48%	0.61%	0.51%	0.94%	0.39%	0.26%	0.87%	0.12%	0.24%	0.63%
	20	1.59%	0.70%	0.66%	0.97%	0.32%	0.23%	0.86%	0.10%	0.15%	0.66%
	30	1.52%	0.68%	0.62%	0.81%	0.24%	0.15%	0.77%	0.12%	0.11%	0.62%
	40	1.48%	0.66%	0.58%	0.76%	0.18%	0.17%	0.69%	0.07%	-0.02%	0.56%
	50	1.54%	0.67%	0.64%	0.87%	0.19%	0.22%	0.69%	0.08%	0.00%	0.59%
OR	10	1.94%	0.55%	0.58%	1.11%	0.25%	0.27%	1.01%	0.12%	0.30%	0.73%
	20	1.96%	0.48%	0.53%	1.00%	0.30%	0.24%	0.89%	0.01%	0.23%	0.67%
	30	1.94%	0.48%	0.55%	0.95%	0.25%	0.25%	0.82%	-0.03%	0.20%	0.64%
	40	1.97%	0.52%	0.50%	0.94%	0.28%	0.25%	0.80%	0.03%	0.24%	0.66%
	50	1.96%	0.50%	0.47%	0.90%	0.23%	0.17%	0.79%	0.01%	0.22%	0.64%
SVM	10	1.85%	0.54%	0.55%	0.85%	-0.02%	0.03%	0.86%	0.00%	-0.14%	0.58%
	20	1.90%	0.56%	0.54%	1.00%	0.10%	0.08%	0.72%	-0.05%	-0.11%	0.58%
	30	1.82%	0.62%	0.57%	1.05%	0.05%	0.17%	0.67%	-0.08%	-0.06%	0.58%
	40	1.84%	0.69%	0.59%	1.08%	0.11%	0.12%	0.77%	0.05%	0.03%	0.64%
	50	1.81%	0.67%	0.58%	1.09%	0.01%	0.16%	0.90%	0.05%	-0.01%	0.64%

The one-way analysis of variance (ANOVA) test was performed to determine whether there was any statistical difference between the means of impact (macro-averaged F-measure of all categories) of a set of independent variables. The independent variables included the feature selection method, topN (i.e., N seedwords for each category as discriminatory concepts for grouping), and other possible factors that were not of interest but can potentially influence the means. The results of ANOVA are listed in Table C.2 in Appendix C. Three-way interaction of $Feature\_Selection*Classifier*Train-TestRatio$ (F-value = 2.28, P-value = 0.0009) was statistically significant at an alpha level of 0.05, and so were three corresponding two-way interactions. As for the two factors of interest, Feature_Selection was statistically significant (F-value = 6.49, P-value <0.0001) while topN was not (F-value = 0.67, P-value = 0.6152).

The ANOVA results indicated that there were significant statistical differences among six feature selection methods. Thus, two post hoc tests of Tukey's and Fisher's LSD were performed for multiple comparison. Both tests showed consistent grouping results. As Table C.3 shows, the mean impact of the CHI feature selection was the highest, followed by OR, LR, LR, MNB, SVM, and MI. Although OR was lower than CHI, they had no significantly statistical difference. However, the mean impact of CHI was still significantly higher than the rest of feature selection methods, including: LR, MNB, SVM, and MI. Thus, the semantic grouping with the seedwords suggested by CHI or OR was statistically more effective in improving the classification performance than LR, MNB, SVM, or MI.

The three-way interaction: Feature_Selection*Classifier*Ratio was examined graphically with three two-way Feature_Selection*Classifier interaction plots at each level of Train-Test Ratios. The three-way interaction plots, Figures 7.22-7.24, graphically showed the effect of semantic tagging with different feature selection methods on the impact of classification performance for MNB, SVM, and LR, with one for each train-test ratio scenario. The impact was measured by the macro-averaged F-measure of all 30 categories. The figures suggest that some feature se-

lection methods were superior in some experiment conditions (i.e., combination of classifier and train-test ratio). However, there was no consistent "best" or "worst" feature selection method for any of the classifiers or train-test ratio scenarios.

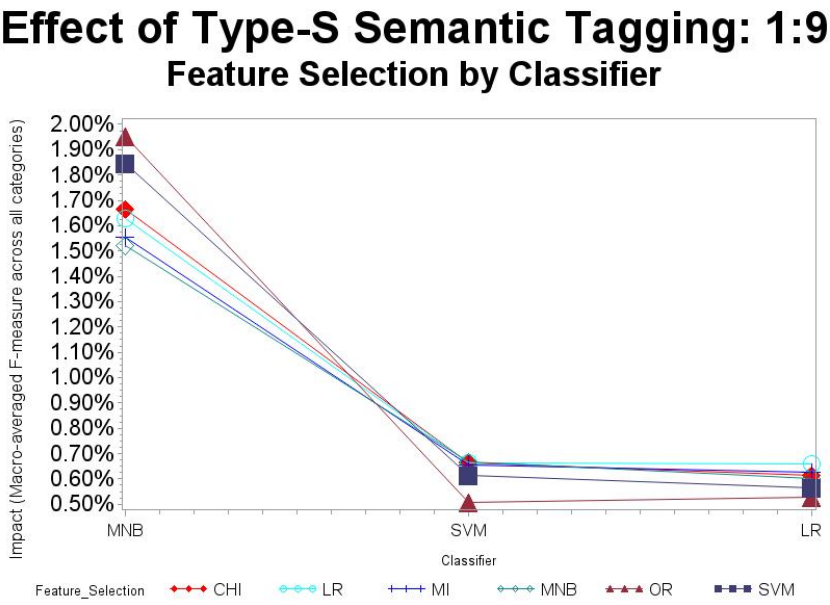

Figure 7.22: Effect of Type-S Semantic Tagging: Feature Selection by Classifier at Train-Test Ratio of 1:9

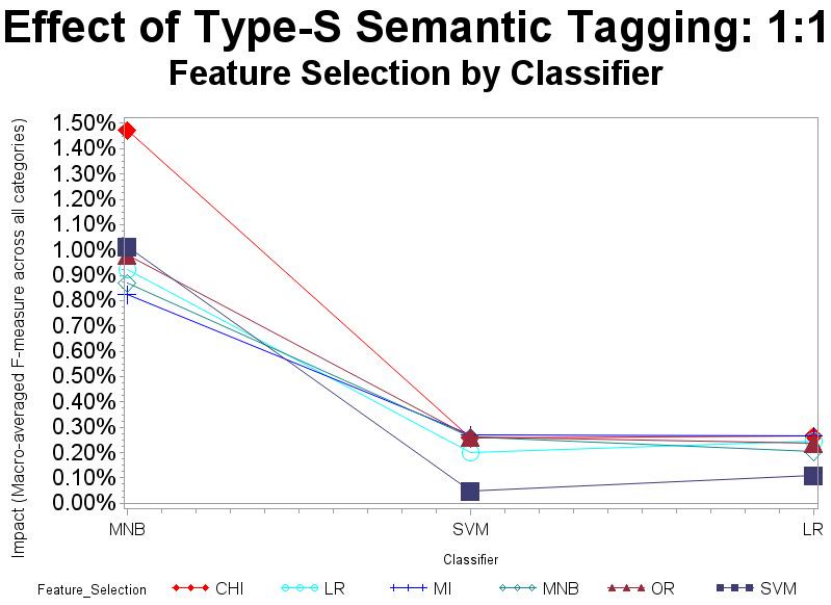

Figure 7.23: Effect of Type-S Semantic Tagging: Feature Selection by Classifier at Train-Test Ratio of 1:1

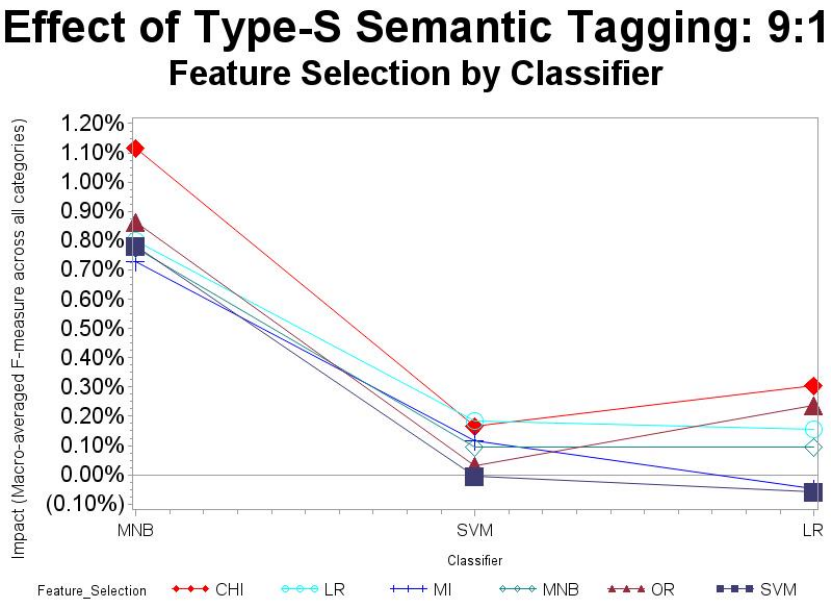

Figure 7.24: Effect of Type-S Semantic Tagging: Feature Selection by Classifier at Train-Test Ratio of 9:1

By ignoring the minimal difference between feature selection methods, the overall semantic tagging effect was examined graphically with the two-way Classifier*Train-Test Ratio interaction plot of Figure 7.25.

Overall, the semantic tagging had a positive impact on improving the classification performance for all classifiers in all three train-test ratio scenarios. The order of impact level, from highest to lowest, is 1:9, 1:1, and 9:1, suggesting that the automated semantic tagging becomes more effective when the availability of training examples decreases. Furthermore, semantic tagging was found to have a greater impact on the classification performance of MNB than LR and SVM. For example, given CHI as the feature selection method, the classification performance of MNB was improved by 1.42% on average (range: 1.02% to 1.72%), compared to 0.39% for LR (range: 0.21% to 0.70%) and 0.36% for SVM (range: 0.08% to 0.69%).

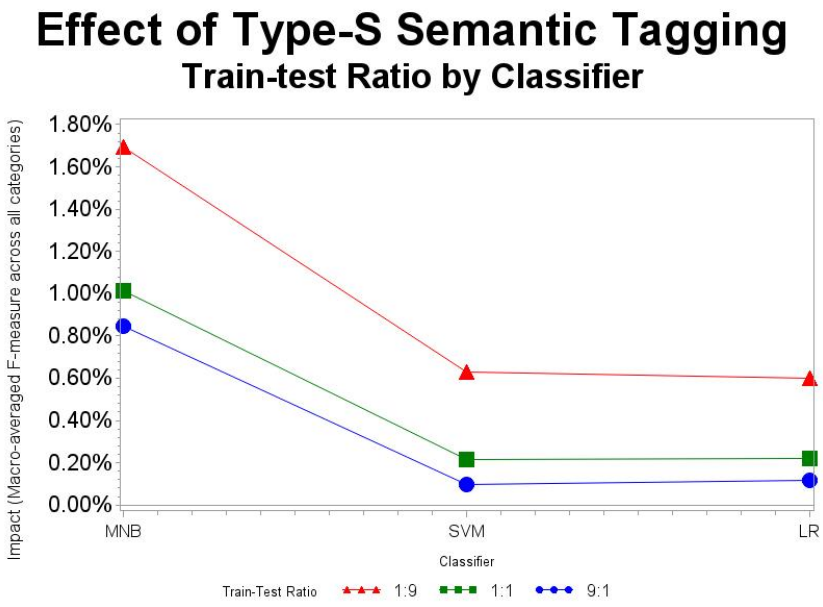

Figure 7.25: Overall Effect of Type-S Semantic Tagging: Train-Test Ratio by Classifier

Next, the effect of automated Type-S Semantic Tagging was evaluated at different levels of category sizes (L, M, S) for MNB, SVM, and LR in three train-test ratio scenarios. Table 7.45 and Figures 7.26-7.28 summarize the overall effect. Overall, Type-S Tagging was most effective in improving the classification performance of small categories (1.04%), followed by the medium categories (0.47%). However, Type-S Tagging was found to slightly decrease the performance of large categories (-0.04%) except when the train-test ratio is 1:9 (+0.02% for SVM and +0.03% for LR).

Table 7.45: Effect of Type-S Tagging on Classification Performance of Three Category Sizes

Category Size	**1:9**			**1:1**			**9:1**			**Overall**
	MNB	**SVM**	**LR**	**MNB**	**SVM**	**LR**	**MNB**	**SVM**	**LR**	
L	-0.02%	0.02%	0.03%	-0.15%	-0.04%	-0.04%	-0.11%	-0.08%	-0.06%	-0.04%
M	1.54%	0.50%	0.51%	0.65%	0.15%	0.24%	0.28%	0.07%	0.00%	0.47%
S	2.24%	0.93%	0.84%	1.78%	0.36%	0.24%	1.89%	0.17%	0.32%	1.04%

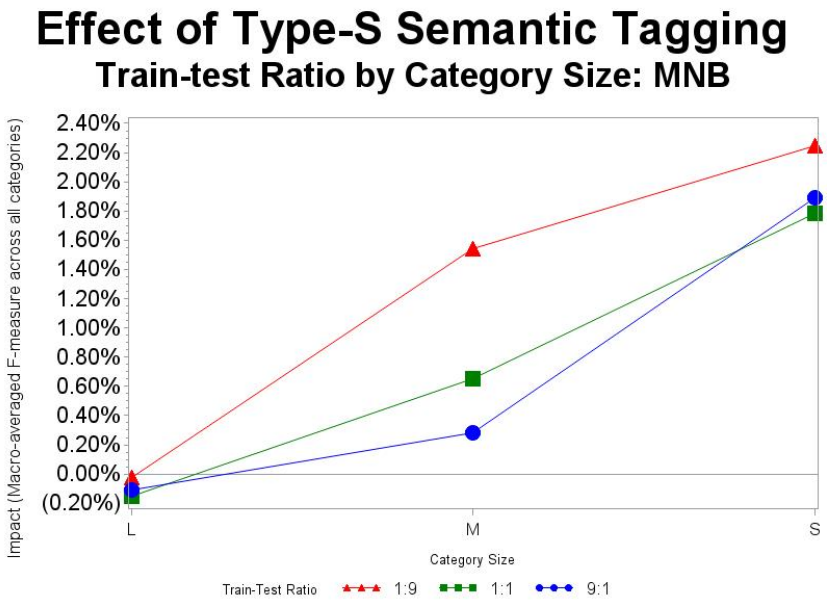

Figure 7.26: Effect of Type-S Semantic Tagging: Train-Test Ratio by Category Size for MNB

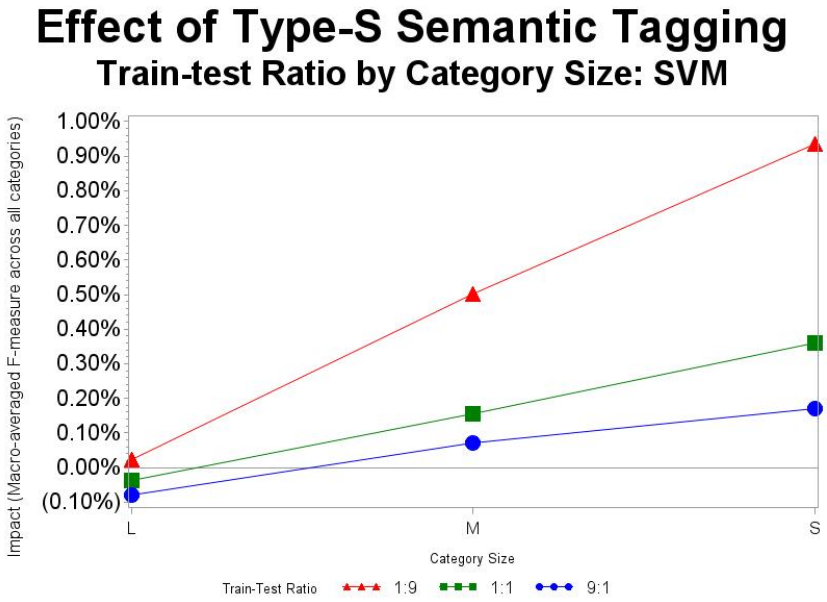

Figure 7.27: Effect of Type-S Semantic Tagging: Train-Test Ratio by Category Size for SVM

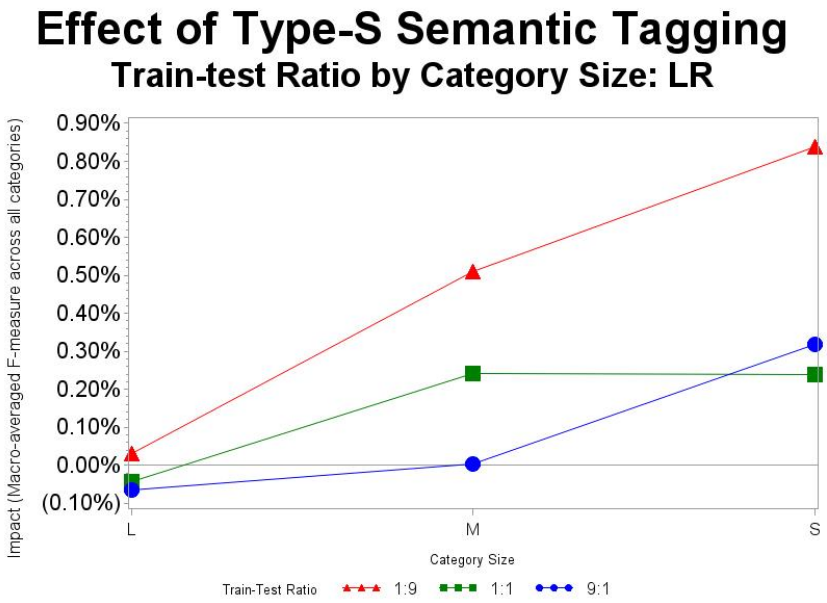

Figure 7.28: Effect of Type-S Semantic Tagging: Train-Test Ratio by Category Size for LR

In summary, Section 7.3 examined the effectiveness of semantic tagging paired with the classic feature selection method to identify discriminatory concepts for grouping for the purpose of improving classification performance. The tested feature selection methods were Chi-square statistic (CHI), Odds ratio (OR), Mutual information (MI), and the coefficient matrices of three classic classifiers (MNB, SVM, LR). The results suggested that CHI and OR were statistically significantly better than the rest of the methods in identifying seedwords as discriminatory concepts for Type-S Tagging. The automated Type-S Semantic Tagging was effective in improving classification performance for MNB, SVM, and LR in three train-test ratio scenarios. The averaged improvement of classification performance, from largest to smallest, was 1.42% for MNB (range: 1.02% to 1.72%), 0.39% for LR (range: 0.21% to 0.70%) and 0.36% for SVM (range: 0.08% to 0.69%). Overall, Type-S Tagging was more effective when the availability of training examples was limited, that is, either the train-test ratio or category size was small. When the

train-test ratio decreased from 9:1 to 1:9, additional positive effect of Type-S Tagging on the classification performance was observed: on average +0.55% for MNB, +0.50% for SVM, and +0.31% for LR. Type-S Tagging had the highest positive impact on the classification performance for small categories (+1.04%) and a moderate positive for medium categories (+0.47%) while a slightly negative for large categories (-0.04%).

8. STUDY SUMMARY AND FINAL EVALUATION

In natural language processing, the so-called "curse of dimensionality" indicates the inherent problem of the sparse, huge Vector Space Model (VSM) to address data rarity in text classification, in addition to its resulting computational complexity. Rare events are the limited labeled training samples associated with the imbalance between classes or training-test data. Classifying small categories in an imbalanced dataset or predicting huge test dataset with small training dataset are considered two of the primary challenges in machine learning. As a result, feature selection or word normalization (grouping) methods are often implemented, as a rule of thumb, in text preprocessing prior to model training. High and low frequency words are often removed from the feature space (i.e., vocabulary) because they are considered to have the least discriminatory power and contribute significantly to the size of vocabulary or total word occurrences in a corpus. However, some studies have demonstrated the improved performance by keeping these extreme-frequency features in statistical text analysis. Due to the lack of research in this field, one object of this study aimed to systematically examine the importance of extreme-frequency words in text classification of injury surveillance data. Chapter 4 was dedicated to achieving this goal and testing Hypotheses 1 to 4. The key results will be highlighted and summarized in Section 8.1.

Another typical task in text preprocessing is word normalization or grouping, which removes the ending of words and merges words with the same root or base form accordingly. Such word grouping methods are instrumental in reducing the volume of low-frequency words and thus the feature space as well. Classic word grouping methods such as stemming and lemmatization assume that words with the same root or base form share the similar semantic meaning, and thus should be grouped and represented with one representative form. Ideal word grouping

can reduce the dimensionality of feature space and increase the density of VSM; hence, it improves not only the computational efficiency, but also statistical robustness of discriminatory concepts and opportunity to train a learning algorithm based on a representative VSM and then make better predictions. However, stemming and lemmatization are not without limitations, for example, their incapability of grouping misspellings, domain-specific words, or words with similar concepts but dissimilar spelling. An ideal word grouping method should be able to group same-hypernym (similar-spelling) words either in similar (such as "electricity" and "electrical") or dissimilar spelling (such as "antidepressant" and "antibiotics" both belong to "drug"). Therefore, the other objective of this study was to propose the so-called Type M+S Grouping Method that can address the identified limitations of classic word grouping methods by considering the linguistic features (morphology and semantics) of words. The proposed grouping method utilized statistical techniques (for identifying predictive categories, selecting discriminatory features, and acquiring semantics) in order to minimize human effort and automate the word grouping process, while allowing for optional incorporation of human input. Chapters 5 to 7 were dedicated to the exploration, introduction and validation of two major elements of the proposed method, Type-M Morphological Mapping and Type-S Semantic Grouping, along with the involved statistical techniques. In Section 8.2, the effectiveness of the Type M+S Grouping Method is examined and Hypotheses 4 to 6 are tested accordingly.

8.1 Role of Extreme-Frequency Words in Text Classification of Injury Data

In Chapter 4, the importance of high- and low-frequency words in text classification was explored. The importance of words was evaluated by the impact on classification performance due to their removal from the feature space. Words were considered important if their absence made a classifier incapable of classifying relevant cases and negatively impacted on classification performance.

Since some high-frequency words are semantically meaningful and indicative of categories, this study focused on the high-frequency words that were stopwords (i.e., functional words that had little semantic meaning but essential to maintain the grammatical relationship with other words). Stopwords were classified into five types: grammatical articles (e.g. *a*, *an*, *the*), pronouns (e.g. *me*, *you*, *she*, *his*), auxiliary verbs (e.g. *was*, *will*, *has*), prepositions (e.g. *by*, *on*, *off*), and others (e.g. *but*, *if*, *or*, *because*). The importance of each stopword type for text classification was examined and the results showed that removing these five types of stopwords had overall positive impact on classification performance. Although preposition-stopwords had the strongest positive impact overall and in many categories, their removal caused a significantly negative impact on classification performance of some other categories. For example, the F-measure of the PEDESTRIAN category decreased by 3.6% on average. Without preposition-stopwords, PEDESTRIAN cases were often misclassified as MOTORVEHICLE. Many PEDESTRIAN cases involved the expressions such as "run over by car", "hit by car", and "struck by car". The presence of the word "by" along with "car" often implied the involvement of an injured person in PEDESTRIAN cases. Since the word "car" served as a discriminative feature for identifying the MOTORVEHICLE category, a PEDESTRIAN case with the word "car" occurring alone without "by" was likely to be mistakenly predicted as MOTORVEHICLE, which was a much larger category that most classifiers were biased in favor of. Consistent with Riloff (1995)'s finding regarding identifying joint-venture activities, both studies indicated that the prepositions were essentially important in building complex phrases to provide specific concept that was often different from the pairing word alone.

Hypothesis 1 of this study was regarding the importance of high-frequency words. Given the results above, Null Hypothesis 1 was falsified and Alternative Hypothesis 1 was accepted because, despite the positive overall effect, the removal of preposition-stopwords negatively impacted the classification performance of certain categories (such as PEDESTRIAN).

While high-frequency words contribute greatly to total word occurrences of a corpus, low-frequency words (LFWs) contribute significantly to the size of vocabulary. The Zipf's Law claimed that, given a corpus of natural language utterances, words that occurred only once take up half of a vocabulary, which was validated in the QISU dataset. Despite that fact that removing low-frequency words seems to become a standard task in text preprocessing, the conflicting empirical evidence and lack of systematic research have kept the question of whether to keep or remove LFWs for text classification unanswered. In Chapter 4, the importance of low-frequency words in text classification was also examined following the investigation of high-frequency words. Given the QISU dataset with a half-million narratives, the LFWs in this study were defined as the words that occurred in less than ten documents in the QISU dataset (i.e., the word occurred once, up to nine times, in every five hundred thousand documents). The macro-averaged F-measure of three classifiers (MNB, SVM, LR) was calculated at each level of Document Frequency Cut-off (DFC), from DFC2 (removing DF1 words) to DFC10 (removing DF1 to DF9 words), in three train-test ratio scenarios (1:9, 1:1, 9:1). Note that "DFx words" stands for the words that occurred in a number of x documents. The overall effect of LFW removal was evaluated by the average of macro-averaged F-measures across DFC2 to DFC10. Overall, removing LFWs had two distinct patterns for three classifiers: as DFC increased, the classification performance was insignificantly decreased for LR (-0.09%) and slightly decreased for SVM (-0.55%) while significantly increased for MNB (+10%). In addition, the overall removal effect also varied with different category sizes and train-test ratios. Removing LFWs tended to have more impact, either positive for MNB or negative for SVM and LR, on small categories and low train-test ratio (1:9) than large categories and high train-test ratio (9:1).

Furthermore, I believed that the frequency of a word could be more fairly defined by considering its sampling size. In an imbalanced dataset, words that have the same occurrence frequency may not have the same probability. Words from

small categories generally have a lower frequency because of small sampling size, and vice versa. LFWs from small categories should not be considered the same as the LFWs from large categories as their sampling sizes are different. In this sense, using the frequency normalized (divided) by category size seems to be a fairer comparison. Thus, the Transformed Frequency (TF) was defined as the sum of the occurrence count normalized by the size of category within which the word occurred, over all the categories that contained the word. For extremely rare words that occurred only once in a dataset, the TF of these DF1 words is simply the inverse of the category probability.

It was of interest to explore the impact of removing extremely LFWs that occurred only once, DF1 LFWs, by TF (i.e., removing DF1 words from the largest to smallest categories) on classification performance for different category sizes in different train-test ratio scenarios. Each classifier showed a distinct pattern of impact, in agreement with the averaged effect of removing DF1-DF9 LFWs. For MNB, removing DF1 LFWs by TF significantly increased the classification performance for small categories and moderately for medium categories while insignificantly for large categories. The positive effect of LFW removal on MNB was the greatest with sufficient training data (at 9:1 train-test ratio). For SVM, removing DF1 LFWs by TF had a positive impact on large categories (greatest at 9:1 ratio) while negative impact on medium and small categories (greatest at 1:9 ratio). For LR, removing DF1 LFWs had no-to-minor negative impact on medium and large categories (within the range of +/- 0.05%) while displaying U-shaped trend of impact (first going negative then turning positive when DF1 LFWs from small categories were starting to be removed) for small categories. Further investigation was made to explain the results by examining the effect of removing DF1 LFWs by TF on classification performance of small categories at the category level. In addition to the overall trends, I focused on the change of classification performance at the TF Cutoff Level right after the DF1 LFWs from the same category were removed. The effects of removing same-category DF1 LFWs were classified into

three types: Effect-D, Effect-I, and Effect-N. Performance decreases (Effect-D) or increases (Effect-I) sharply after same-category DF1 LFWs are removed and stays at similar level afterwards, or it fluctuates in no-or-limited association (Effect-N) with the removal of DF1 LFWs. The results suggested that the occurrence of Effect-D was independent of category sizes or train-test ratios. Effect-N was more likely to occur when training data were sufficient (large categories or high train-test ratio) while Effect-I was more likely to occur when training data were limited (small categories or low train-test ratio). Effect-I and the U-shaped trend of LFWs removal impact observed in small categories for LR can be considered as an artifact of overfitting as LR is prone to overfitting in the case of limited training examples.

Given these results, Hypotheses 2 to 4 were tested accordingly. Null Hypothesis 2 was falsified for SVM and LR, but not for MNB as removing LFWs negatively impacted the classification performance of SVM and LR, while significantly improving MNB. Thus, Alternative Hypothesis 2 was rejected for MNB but accepted for SVM and LR. Since Null Hypothesis 2 was accepted for MNB, Hypotheses 3 and 4 were not tested for MNB. Null Hypothesis 3 was falsified and Alternative Hypothesis 3 was accepted for SVM. However, Null Hypothesis 3 was not falsified and Alternative Hypothesis 3 was not accepted for LR because removing LFWs had a positive impact on small categories by alleviating the overfitting problem. Without these extremely low-frequency words in the model, LR is able to assign weights more evenly and properly. As a result, the model can rely on more frequent and discriminatory features, rather than the rare features that may never occur again in a test set. Null Hypothesis 4 was falsified and Alternative Hypothesis 4 was accepted for SVM and LR because removing LFWs had more severe negative impact at the train-test ratio of 1:9 than 9:1.

8.2 Utilization of Low-frequency Words for Improving Text Classification of Injury Data

8.2.1 Summary of Type-M Mapping and Type-S Grouping

To reduce the size of feature space and improve the statistical robustness of discriminatory concepts in text classification, the classic word normalization (grouping) methods of stemming and lemmatization are often applied to group same-hypernym (similar-concept) words with similar spelling through mapping (replacing words with their root or base form). As noted in Chapter 5, these classic methods have some limitations in the following:

- Incapability of handling misspellings, domain-specific words (such as names of drugs or chemicals), and same-hypernym
- Arguable assumption that words with the root or base form have similar meaning and do not predict a conflicting category
- Exclusion of rare features that occurred only in test set but not training set

As a way for utilizing and reducing rare words (seen or unseen in training set), the proposed Type M+S Grouping Method aimed to address the above limitations of stemming and lemmatization by grouping words based on their linguistic features, morphologically and semantically. The Type M+S Grouping Method has two parts: "Type-M Morphological Mapping" for grouping same-hypernym words with similar spelling and "Type-S Semantic Grouping" for grouping same-hypernym words with dissimilar spelling.

In Chapter 6, the Type-M Morphological Mapping Method was introduced to group same-hypernym words with similar spelling (morphology). Type-M Mapping is similar to stemming and lemmatization because they all merge words based on their spelling, but the proposed method is less aggressive and more grounded by considering the predictive category of words and only grouping similar-spelling

words when they predict a non-conflicting category. In order to do that, Type-M Mapping requires the measure of morphological similarity and predictive strength for categories. This study used the n-gram similarity module in Python to algorithmically quantify morphological similarity between words and used the coefficient matrix of a linear classifier to signify the strength of words for predicting categories and thus to determine the predictive category for words. The essence of Type-M Mapping is to unconditionally group words with high n-gram similarity (extremely similar spelling) and then group words with moderate n-gram similarity if they do not predict a conflicting category. The effectiveness of Type-M Mapping in improving classification performance was evaluated based on the impact of classification performance (macro-averaged F-measure of all categories) of three classic classifiers (MNB, SVM, LR) in three train-test ratio scenarios (1:9, 1:1, 9:1). The feasibility of using the coefficient matrix of three classic classifiers (MNB, SVM, LR) to identify predictive categories for words is related to how much Type-M Mapping improved the classification performance. The results of ANOVA test indicated the three classifiers were not statistically different in terms of being an indicator of predictive categories, although Type-M Mapping paired with MNB was found to have the least positive effect on classification performance. For overall effect, Type-M Mapping was able to improve the classification performance by 1.34%, averaged across three classifiers and three train-test ratio scenarios. Type-M Mapping had the greatest positive impact on classification performance for MNB (3.07%), followed by LR (0.32%) and SVM (0.14%). As expected, Type-M Mapping was the most effective in small samples, either in small categories (+2.3% for small categories vs. -0.15% for large categories) or small train-test ratio (1.6% for 1:9 vs. 1.18% for 9:1).

In Chapter 7, the Type-S Semantic Grouping Method was proposed to group same-hypernym words with dissimilar spelling. Type-S Grouping differs from classic word normalization methods and Type-M Mapping by considering semantic meaning of words. To avoid the human effort of assigning semantics manually,

the feasibility of two types of statistical semantics in quantifying semantic similarity between words and identifying same-hypernym words was first explored. "Correlational similarity" concerns the co-occurrence of words, which can be measured by the Pointwise Mutual Information (PMI) that examines the likelihood that two-words tend to co-occur versus occurring alone. On the other hand, "distributional similarity" concerns the context, which can be measured by the cosine of context vectors generated by "Word2Vec", one of the most promising predictive-based distributional semantics models that derives a lower-dimensioned, optimized Vector Space Model based on self-annotated text data. A semantic data mining method was proposed to utilize the statistical similarity measure to identify words with the same hypernym with a given "seedword" (i.e., the word that carries discriminatory concept and is used to calculate the Word2Vec similarity score with other words). The distributional semantics measured by Word2Vec was found to be more effective than the correlational semantics measured by PMI, in identifying words with similar concept (drug in the experimented task). Thus, by falsifying Null Hypothesis 5, Alternative Hypothesis 5 was accepted that distributional semantics is more effective than correlational semantics in identifying same-hypernym words. Consequently, Word2Vec was used as a statistical semantic measure of quantifying word similarity for the proposed Type-S Grouping Method. In addition, an exploratory study for Word2Vec was conducted to gain insight into the mechanism of how Word2Vec quantifies word similarity and ranks similar words. The results suggested that Word2Vec tended to give high weight on rare features, which could be an advantage (effectively identify morphological variants or misspellings of important concepts) and also a disadvantage (noisy results with high similarity with some random, irrelevant low-frequency words). As a remedy, setting higher threshold (i.e., lowest Word2Vec similarity for a word to be considered as a same-hypernym candidate) or manually reviewing the candidate word list suggested by Word2Vec can filter out noises from final results of same-hypernym words.

Given a set of seedwords that carry discriminatory concepts, the Type-S Semantic Grouping Method was proposed to group words, identified by Word2Vec, semantically similar to pre-selected seedwords. To evaluate the effectiveness of Type-S Grouping, five types of poisoning and allergy injury agents ("PA-concept": drug, alcohol, chemical, food, plant) were selected for testing. The effect of Type-S Grouping for words with PA-Concept on classification performance of associated categories was examined for three classifiers in three train-test ratios. In the experiment, three factors were tested, including two grouping strategies (mapping and tagging), two threshold ("low": 0.2 and "moderate": 0.4), and three review levels (no-review, one-time review, ultimate review). Tagging was found superior and thus used for the following experiments of semantic grouping (called "Type-S Tagging"). The results implied that manual review was inevitable with a low or moderate threshold (0.2 to 0.4) unless Type-S Tagging was applied at small train-test ratio (1:9). Also, one-time review was found to have the best cost-benefit ratio and thus more practical than ultimate review.

To improve the scalability of Type-S Tagging, this study also examined the feasibility of using classic feature selection methods to automatically identify seedwords that indicate discriminatory concepts for grouping. Using Type-S Tagging paired with Chi-square statistics and Odds Ratio were statistically significantly more effective in improving classification performance than pairing with the other feature selection methods (Mutual Information and the coefficient matrix of MNB, SVM, and LR). The automated Type-S Tagging Method was able to improve the classification performance (macro-averaged F-measure across 30 categories) by 0.76% on average, with more positive impact on MNB (1.42%) than LR (0.39%) and SVM (0.36%). Similar to Type-M Mapping, Type-S Tagging was more effective with in small training samples, either small train-test ratio (1% for 1:9 vs. 0.53% for 9:1) or small category size (+1.04% for small categories vs. -0.04% for large categories).

8.2.2 Evaluation of Type M+S Grouping and Add-on Methods

With combined Type-M Morphological Mapping and Type-S Semantic Tagging, the Type M+S Grouping Method was evaluated and benchmarked against two classic word normalization methods: stemming and lemmatization. Tables 8.1 and 8.2 list the numbers of morphs/tags and words being mapped/tagged for Type-M Mapping and Type-S Tagging in three train-test ratio scenarios.

Table 8.1: Type-M Mapping: Numbers of Words Being Mapped and Morphs

Train-Test Ratio	# Words Being Mapped	# Morphs
1:9	13,065	4,480
1:1	14,511	6,031
9:1	13,343	6,321

Table 8.2: Type-S Tagging: Numbers of Words Being Tagged and Tags

Train-Test Ratio	# Words Being Tagged	# Tags
1:9	886	161
1:1	1,271	264
9:1	1,340	241

Three add-on methods that potentially can further improve the automated Type M+S Grouping Method were tested, which are:

- R**eviewed Tagging for PA concept ("PA")**: Manually review the Word2Vec-suggested word candidates (13% of the vocabulary, 6221 words) to classify them into five types of poisoning and allergy-related injury agents (i.e., drug, chemical, alcohol, food, plant) and then tag these 892 word candidates with

five types of manually-verified labels. Refer to Section 7.2.2 for the Semantic Tagging Method with a threshold of 0.4 and one-time review.

- **Two-word sequence tagging ("S2")**: Tagging the top 30 most discriminative two-word sequences identified by the feature selection method of Chi-square statistic for the categories of BICYCLE, MOTORBIKE, MOTORVEHICLE, and PEDESTRIAN. Initial studies suggested that not all of the categories benefited from two-word sequence tagging. As a result, only the above four categories were considered and the discriminatory two-word sequences for each category were prioritized using the Chi-square statistic, the best-performing feature selection method identified in Section 7.3.

- **Naive Bayes log-count ratio as input features for SVM and LR Classifiers ("NB")**: Applying NB-SVM or NB-LR to classify processed injury narratives after grouping. Wang and Manning (2012) presented a simple model variant where a SVM was developed with the NB log-count ratio as feature values of Vector Space Model. To test the feasibility of pairing NB-SVM and NB-LR with the proposed grouping method in improving the classification performance, this study developed NB-LR by modifying the Rei (2015)'s implementation of NB-SVM in Python. NB-weighted classifiers are built on a weighted Vector Space Model and tend to give a higher weight to low-frequency features; thus, NB-SVM and NB-LR have the potential to help classification performance of small categories, which are verified in this study.

Table 8.3 lists eight grouping models, which are combinations of the five proposed methods: Type-M Mapping ("M"), Type-S Tagging ("S"), and three add-on methods ("PA", "S2", "NB"). Note the third add-on method "NB" only applies to SVM and LR. Table 8.4 lists the applicability of the eight proposed grouping models for MNB, SVM, and LR.

Table 8.3: Word Grouping Models: Combinations of Purposed Methods

	Grouping Model: Combination of Proposed Methods							
Proposed Method	**MS**	**MSPA**	**MSS2**	**MSS2PA**	**MS_NB**	**MSPA_NB**	**MSS2_NB**	**MSS2PA_NB**
M: Type-M Morphological Mapping for same-hypernym words with similar spelling	x	x	x	x	x	x	x	x
S: Type-S Semantic Tagging for same-hypernym words with dissimilar spelling	x	x	x	x	x	x	x	x
PA: Reviewed Tagging for PA concept		x		x		x		x
S2: Two-word sequences tagging for categories BICYCLE, MOTORBIKE, MOTORVEHICLE, PEDESTRIAN			x	x			x	x
NB: Naive Bayes log-count ratio as input features					x	x	x	x

Table 8.4: Applicability of Proposed Word Grouping Models for MNB, SVM, and LR

	Grouping Model: Combination of Proposed Methods							
Applicability for Classifier	**MS**	**MSPA**	**MSS2**	**MSS2PA**	**MS_NB**	**MSPA_NB**	**MSS2_NB**	**MSS2PA_NB**
Multinomial Naive Bayes (MNB)	x	x	x	x				
Support Vector Machine (SVM)	x	x	x	x	x	x	x	x
Logistic Regression (LR)	x	x	x	x	x	x	x	x

The effectiveness of proposed grouping models was benchmarked with classic grouping methods, stemming ("STEM_DFC1") and lemmatization ("LEMMA_DFC1"). Table 8.5 tabulates the impact of proposed word grouping models and benchmarked methods on the classification performance (macro-averaged F-measure of all categories). The impact was measured by the performance difference between grouping and non-grouping (i.e., standard classification with feature space of entire vocabulary). Figure 8.1 graphically shows the results of impact presented in Table 8.5. Figure 8.2 visually displays the final classification performance of grouping methods to compare against the standard classification (annotated with "control"). Table D.1 in Appendix D tabulates the final macro-averaged F-measures of three classifiers for overall categories and three category sizes for all grouping methods and standard classification. Although this study focused on the grouping effect at the level of category sizes rather than categories, the category-wise results for each classifier in each train-test ratio scenario are provided in Tables D.2-D.10 for reference.

Table 8.5: Benchmarking for Proposed Word Grouping Models for MNB, SVM, LR

Grouping Method	MNB	SVM	LR	Overall
STEM_DFC1	1.58%	0.29%	0.63%	0.83%
LEMMA_DFC1	0.35%	0.28%	0.49%	0.37%
MS	3.55%	0.36%	0.65%	1.52%
MSS2	4.20%	0.53%	0.88%	1.87%
MSPA	3.72%	1.03%	1.99%	2.25%
MSS2PA	4.39%	1.22%	2.20%	2.60%
MS_NB		-1.85%	1.21%	-0.32%
MSS2_NB		-1.72%	1.37%	-0.18%
MSPA_NB		-0.59%	2.51%	0.96%
MSS2PA_NB		-0.55%	2.69%	1.07%

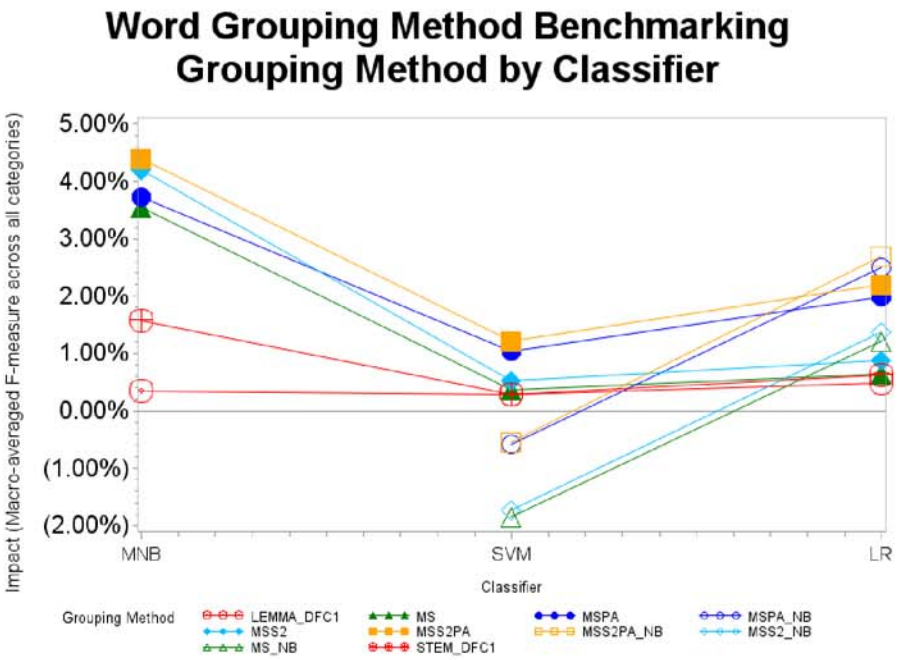

Figure 8.1: Word Grouping Method Benchmarking

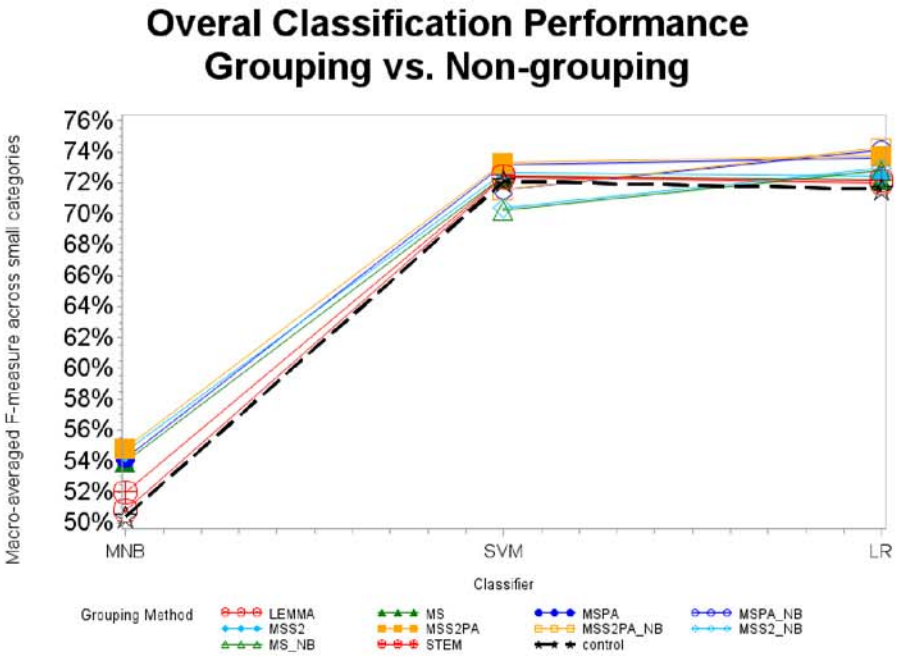

Figure 8.2: Overall Classification Performance: Grouping vs. Non-grouping

Overall, MSS2PA was the most effective grouping method in improving classification performance for SVM (1.22%) and MNB (4.39%) across three train-test ratio scenarios. MSS2PA comprises the automated Type M+S Grouping Method plus two add-on methods: "S2" for two-word sequence tagging and "PA" for reviewed tagging of PA-concept. However, the add-on method of "NB" for supplying Naive Bayes Weighted feature values was able to further improve LR from 71.6% to 74.3%, with an increment of 0.5% in addition to 2.2% from MSS2PA. As a result, the MSS2PA Grouping Model paired with the NB-LR classifier was the best combination for classifying the QISU dataset, improving the overall classification performance of small categories by 4.7%, medium categories by 2.4%, and large categories by 1.4%.

In addition to the positive effect of Type M+S Grouping (MNB: 3.55%; SVM: 0.36%; LR: 0.65%), the proposed add-on methods were able to make additional improvement to the overall performance. Table 8.6 tabulates the effect of add-on methods on improving Type M+S Grouping for overall classification performance, showing all positives except for the case where the "NB" add-on was applied to SVM.

Table 8.6: Impacts of Add-on Methods to Type M+S Grouping Method on Macro-averaged F-measure

Add-on to MS	**MNB**	**SVM**	**LR**	**Overall**
S2	0.65%	0.17%	0.23%	0.35%
PA	0.17%	0.67%	1.34%	0.73%
S2+PA	0.84%	0.86%	1.55%	1.08%
NB		-2.21%	0.56%	
S2+NB		-2.08%	0.72%	
PA+NB		-0.95%	1.86%	
S2+PA+NB		-0.91%	2.04%	

To compare the effect of proposed grouping models and benchmarked methods, the analysis of variance (ANOVA) test was applied to examine whether they were statistically different in their averaged impact. Prior to the ANOVA test that is used, the Levene's Test was conducted first to test the ANOVA assumption of homogeneity of variance. As Table D.11 shows, this assumption was rejected (F-value = 6.97, P-value <0.0001). Despite unequal population variances, the ANOVA test is still considered robust when the populations have equal sample sizes (Montgomery, 2009, pp. 78; Northwestern Medicine, 1997; SAS, n.d.). Thus, the ANOVA test was still performed. The results, listed in Table D.11, indicated that there were statistical differences among these grouping methods (F-value = 16.63, P-value <0.0001). Since the two-way Classifier*Grouping_Method interaction was significant (F-value = 10.04, P-value <0.0001), conducting the post hoc test for multiple comparisons could have misleading results since the comparing the impact means of Grouping Method might be obscured by its interaction with Classifier (Montgomery, 2009, pp. 173). One approach suggested by Montgomery (2009) is to apply the multiple comparison method to impact means of Grouping Method at each level of Classifier (i.e., MNB, SVM, LR). Thus, the three post hoc Tukey's and Fisher's LSD Tests were conducted separately for MNB, SVM and LR. The results of multiple comparisons, listed in Tables D.12-D.14, confirmed that the proposed MSS2PA method was statistically significantly superior to stemming and lemmatization in improving the overall classification performance for MNB, SVM and LR. Also, the grouping method of MSS2PA and MSPA paired with the "NB" add-on (i.e., MSS2PA_NB, MSPA_NB,) was also statistically superior to stemming and lemmatization for LR, even though not for SVM.

Next, it was of interest to compare the effect of keeping ("DFC1" — keeping all words) and removing DF1 Words ("DFC2" — removing words occurred only once) after applying the word grouping methods. "DFC" stands for Document Frequency Cut-off. Figure 8.3 graphically shows the impact of grouping methods on classification performance at DFC1 and DFC2. It can be observed that removing

DF1 words after word grouping showed an insignificant impact on the classification performance for SVM (-0.1%) and LR (0.03%) while a major improvement for MNB (5%).

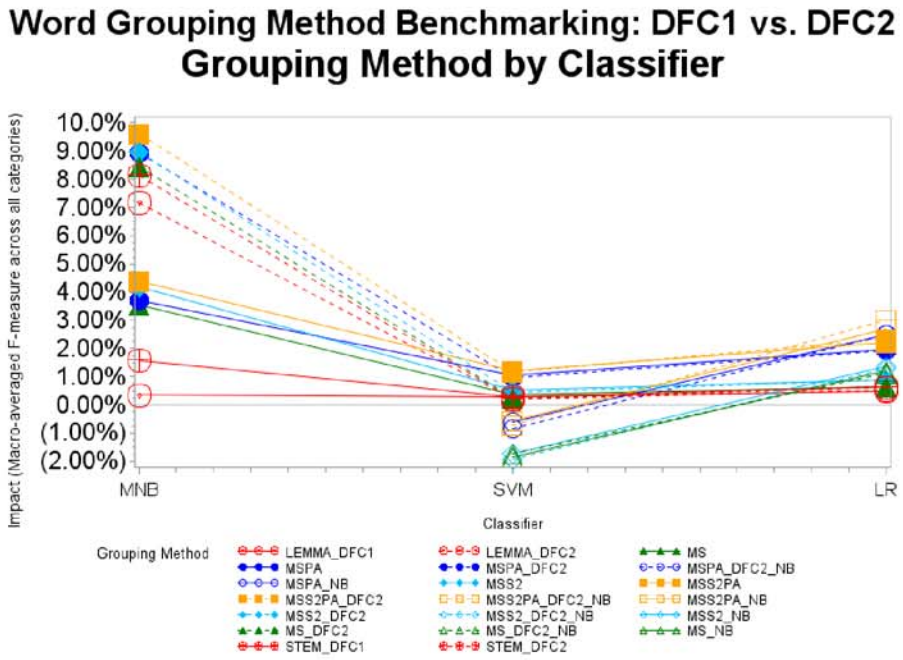

Figure 8.3: Word Grouping Method Benchmarking – DFC1 vs. DFC2

The effectiveness of proposed Grouping Models and benchmarked methods in improving classification performance was also examined at different levels of data rarity, three train-test ratio scenarios (1:9, 1:1, 9:1) and three category sizes (S, M, L). Due to their distinct patterns of impact, three classifiers (MNB, SVM, LR) were examined separately.

First, the proposed Grouping Models are benchmarked with stemming and lemmatization on classification performance for three classifiers in three train-test ratio scenarios. Table 8.7 tabulates and the following Figures 8.4-8.6 visually display the results of benchmarking in three train-test ratio scenarios, with one figure for each classifier. Figures 8.4-8.6 show consistent results with Figure 8.3 while at train-test ratio levels. It can be observed that word grouping methods had

stronger effect when training data were limited at the train-test ratio of 1:9 than at 9:1. As the train-test ratio decreases, the difference between the proposed Grouping Models and benchmarked methods became more significant in terms of the effectiveness of these grouping methods in improving classification performance. Classic grouping methods showed relatively stable effect across three train-test ratio scenarios. In addition, the proposed Grouping Models demonstrated an increasing positive effect as the train-test ratio decreases, which implied that the effectiveness of the proposed methods in utilizing unseen features that occurred only in test dataset, which were not considered in stemming and lemmatization. In addition, the positive effect of applying NB-LR was found to decrease as the train-test ratio increases, suggesting that NB-LR is more effective with the limited training data.

Table 8.7: Benchmarking of Proposed Word Grouping Models (DFC1)

	1:9			1:1			9:1			
Grouping Methods	**MNB**	**SVM**	**LR**	**MNB**	**SVM**	**LR**	**MNB**	**SVM**	**LR**	**Overall**
STEM_DFC1	2.26%	0.57%	0.93%	1.37%	0.16%	0.59%	1.09%	0.13%	0.37%	0.83%
LEMMA_DFC1	0.65%	0.47%	0.49%	0.30%	0.22%	0.39%	0.10%	0.16%	0.60%	0.37%
MS	3.59%	0.89%	1.25%	3.79%	0.16%	0.29%	3.28%	0.03%	0.39%	1.52%
MSS2	5.06%	1.01%	1.51%	4.19%	0.60%	0.73%	3.36%	-0.02%	0.41%	1.87%
MSPA	3.76%	1.91%	3.10%	3.93%	0.74%	1.49%	3.47%	0.44%	1.38%	2.25%
MSS2PA	5.07%	2.00%	3.28%	4.31%	1.24%	1.90%	3.78%	0.41%	1.44%	2.60%
MS_NB	.	-1.07%	2.79%	.	-2.61%	0.67%	.	-1.87%	0.17%	-0.32%
MSS2_NB	.	-1.36%	2.92%	.	-2.41%	1.01%	.	-1.40%	0.17%	-0.18%
MSPA_NB	.	0.89%	4.56%	.	-1.65%	1.98%	.	-1.00%	0.97%	0.96%
MSS2PA_NB	.	0.63%	4.65%	.	-1.33%	2.36%	.	-0.95%	1.07%	1.07%

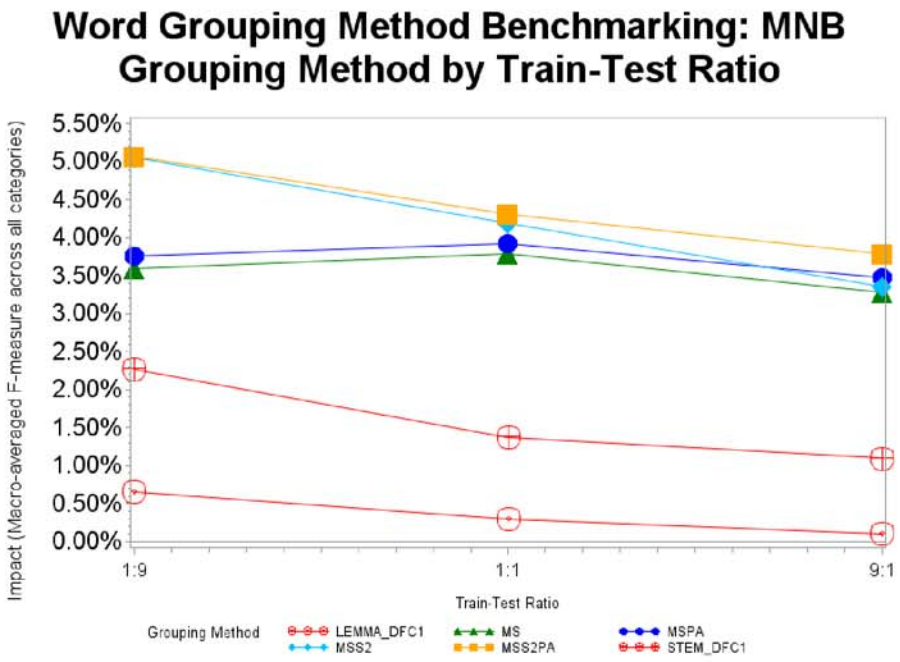

Figure 8.4: Word Grouping Method Benchmarking by Train-Test Ratio for MNB

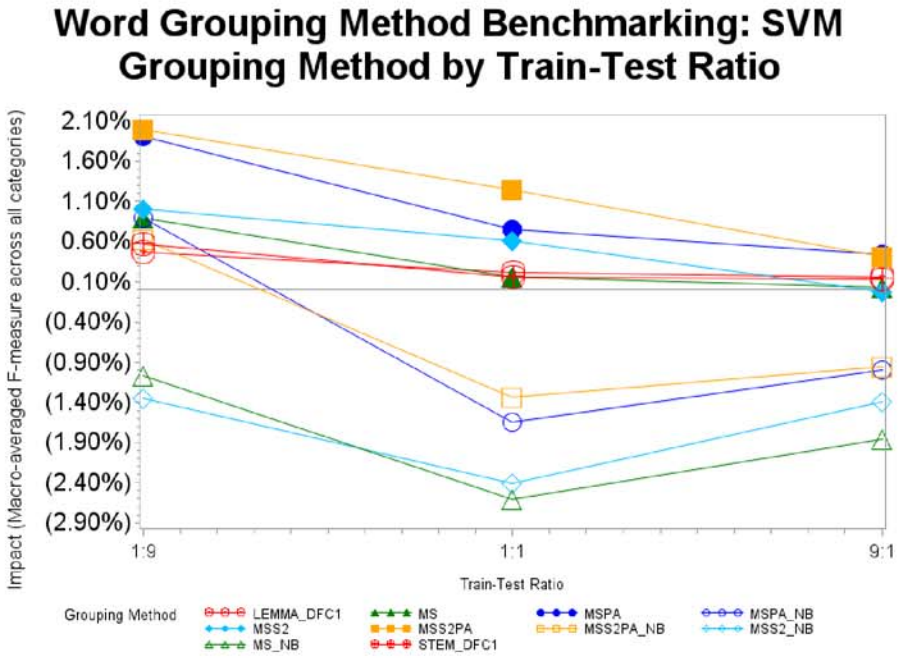

Figure 8.5: Word Grouping Method Benchmarking by Train-Test Ratio for SVM

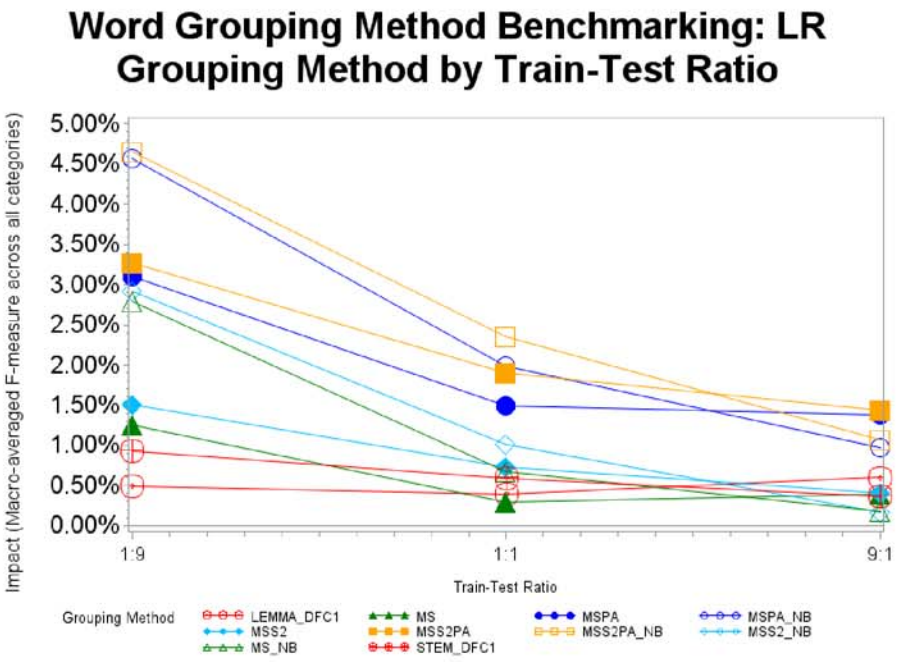

Figure 8.6: Word Grouping Method Benchmarking by Train-Test Ratio for LR

Figures similar to Figure 8.4-8.6 were organized to benchmark the proposed Word Grouping Models with stemming and lemmatization on different category sizes. Figures 8.7-8.9 show the impact of the proposed and benchmarked methods on the classification performance of different-size categories for MNB, SVM, and LR. Overall, the word grouping method had the greatest positive impact on classification performance for small categories, followed by medium and large categories. Similar to the previous finding, classic grouping methods had a relatively flatter effect while the proposed Grouping Models were the most effective when training examples were limited (small categories) rather than sufficient (large categories). The results suggested that, despite small training samples at first, the classification performance of small categories can be improved, to a much greater extent than large categories, by creating better representation of Vector Space Model with more representative and statistically robust (stand-out) features through proper word grouping.

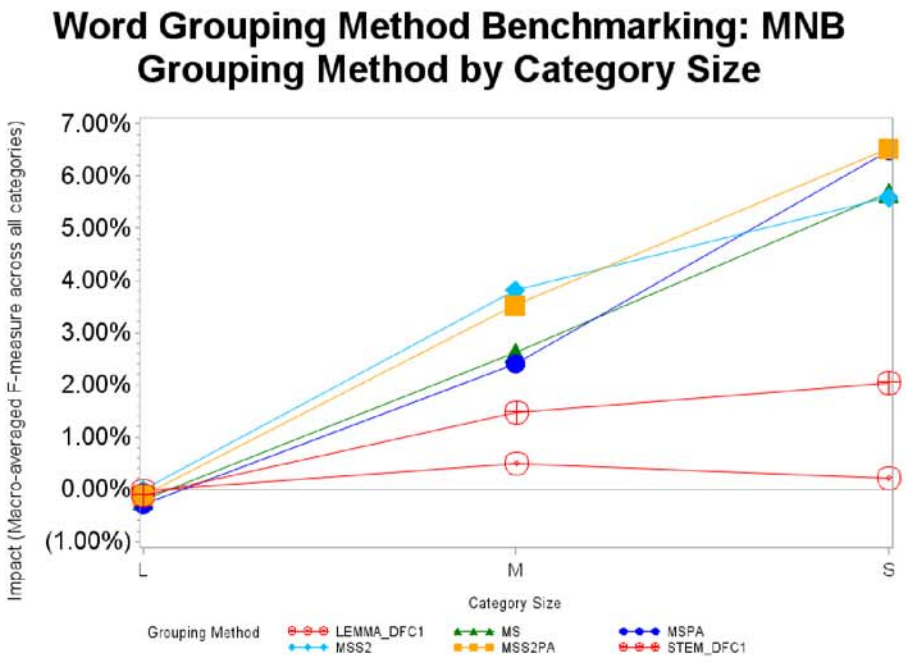

Figure 8.7: Word Grouping Method Benchmarking by Category Size for MNB

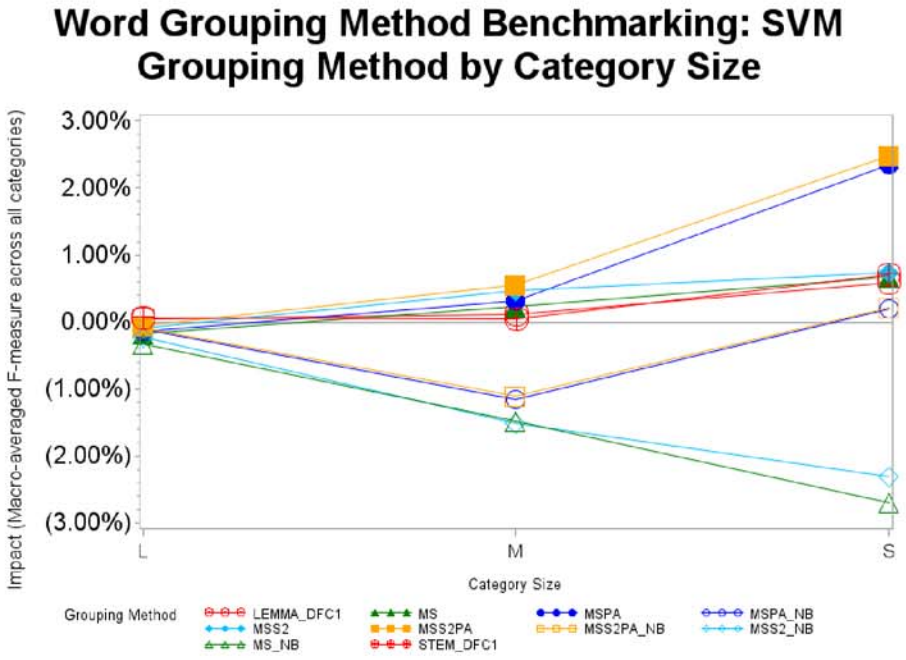

Figure 8.8: Word Grouping Method Benchmarking by Category Size for SVM

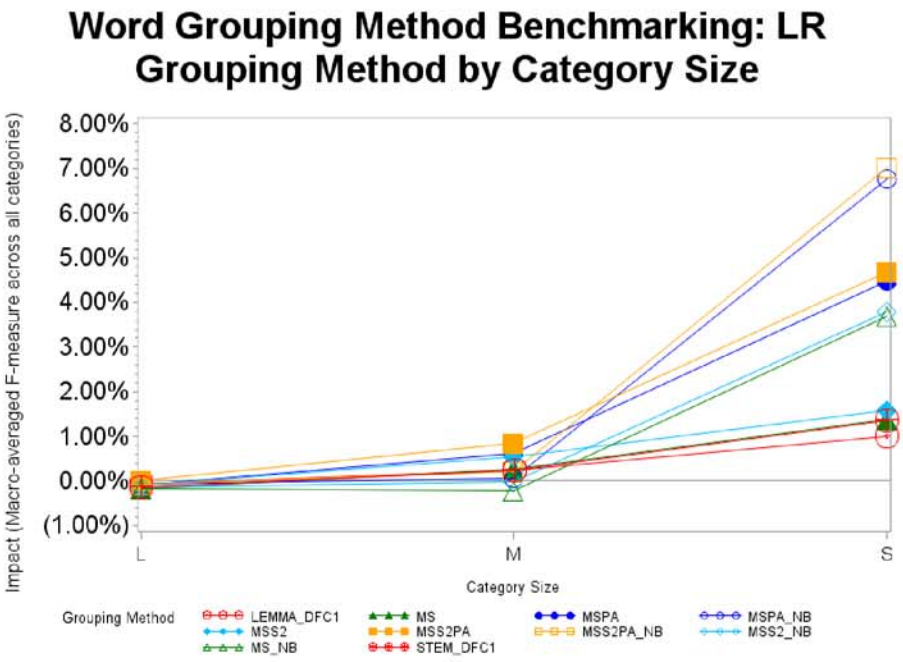

Figure 8.9: Word Grouping Method Benchmarking by Category Size for LR

Until now, the effect of word grouping is mostly represented by impact in terms of a percent change in the macro-averaged F-measure. Figure8.10 shows macro-averaged F-measures of standard classification without grouping ("control," represented with dash lines) and with grouping methods, on three category sizes for three classifiers. Each plot in Figure8.10 represents the relationship between the macro-average F-measure and train-test ratio for a combined setting of category size and classifier. An asymptotic relationship between the classification performance and train-test ratio can be observed: the slope decreases as the train-test ratio increases. The best-performing word grouping ("MSS2PA" for MNB and SVM and "MSS2PA_NB" for LR) had an insignificant impact on the performance of large categories for all three classifiers (0.1%) while improving the performance of medium categories minimally for SVM (0.5%; 1:9: 1%, 1:1: 0.6%, 9:1: 0.1%), slightly for LR (0.9%; 1:9: 1.5%, 1:1: 0.8%, 9:1: 0.3%), and moderately for MNB (3.8%; 1:9: 5.9%, 1:1: 3.4%, 9:1: 2.1%). The word grouping was found to be the

most effective in improving the performance of small categories with the highest impact on LR (7%; 1:9: 11.2%, 1:1: 6.1%, 9:1: 3.7%), followed by MNB (6.6%; 1:9: 5.5%, 1:1: 6.9%, 9:1: 7.5%) and SVM (2.6%; 1:9: 4%, 1:1: 2.5%, 9:1: 1.2%).

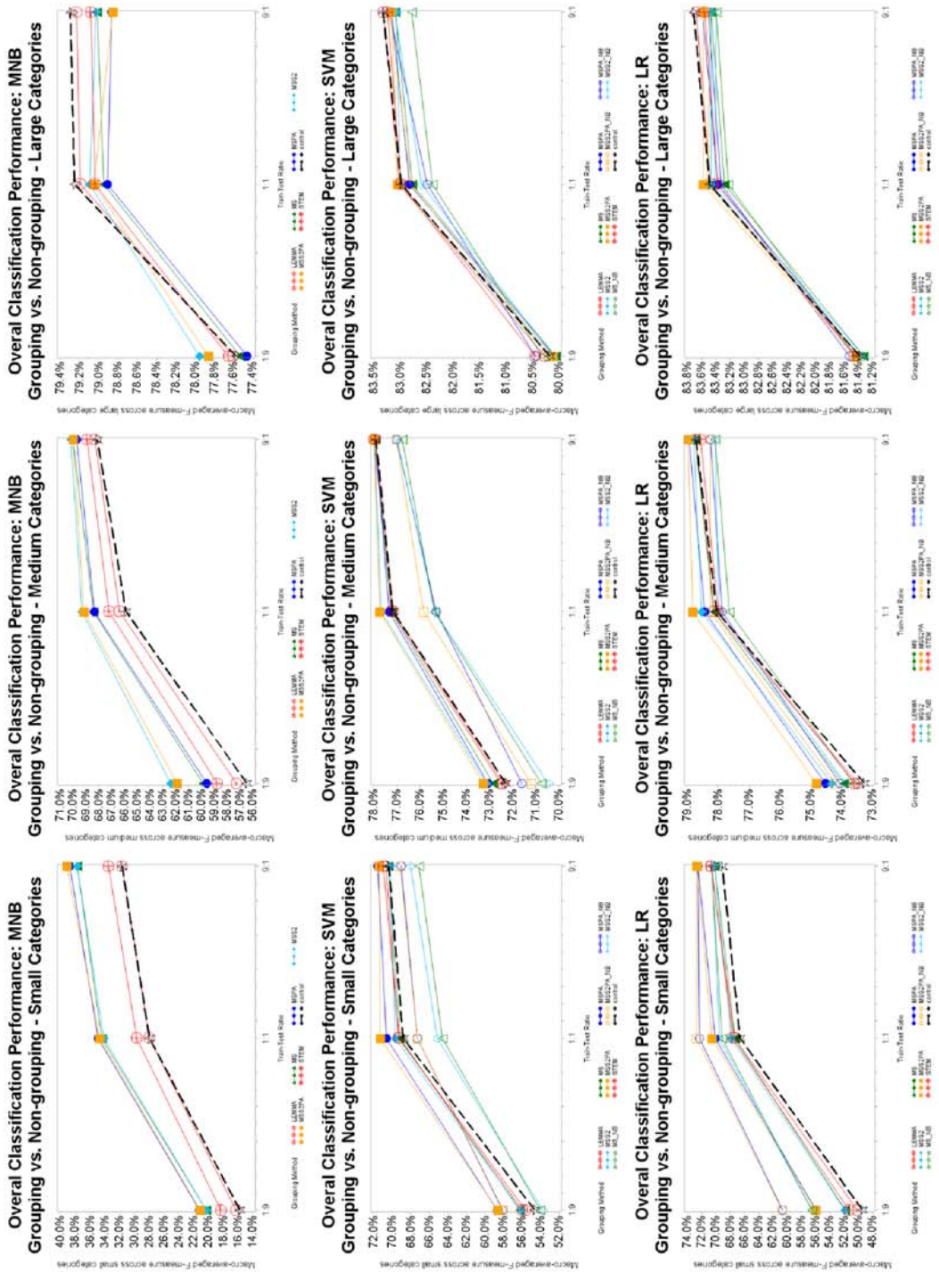

Figure 8.10: Classification Performance on Three Category Sizes for MNB, SVM, LR: Grouping vs. Non-grouping

As Figure 8.10 shows, this asymptotic relationship between the performance and train-test ratio can be approximated with a piecewise defined function that includes two linear equations with a different part of the domains. The two linear equations pass through three data points of (r, F), where r represents the train-test ratio $\epsilon(\frac{1}{9}, 1, 9)$ and F represents the corresponding performance in the macro-averaged F-measure.

By assuming the final performance is a piecewise defined function of train-test ratio, each combined setting of classifier and category size has two linear equations relating macro-averaged F-measure F_i and train-test ratio $r_i \epsilon$ $(\frac{1}{9}, 1, 9)$ for i=1,2,3. One linear equation connects $(\frac{1}{9}, F_1)$ and $(1, F_2)$ and the other connect $(1, F_2)$ and $(9, F_3)$. Under this assumption, any one of the two linear function can be expressed as in the slope-intercept form of:

$$F(r) = \begin{cases} slope_1 \times r + intercept_1; & \frac{1}{9} \leq r < 1 \\ slope_2 \times r + intercept_2; & 1 \leq r < 9 \end{cases} \tag{8.1}$$

The slope and intercept can be derived using two-point formula. Given three points $(\frac{1}{9}, F_1)$, $(1, F_2)$, and $(9, F_3)$:

$$slope_1 = \frac{F_2 - F_1}{1 - \frac{1}{9}} \tag{8.2}$$

$$intercept_1 = F_1 - slope_1 \times \frac{1}{9} \tag{8.3}$$

$$slope_2 = \frac{F_3 - F_2}{9 - 1} \tag{8.4}$$

$$intercept_2 = F_2 - slope_2 \times 1 \tag{8.5}$$

Table 8.8 lists the slope and interception for the two linear functions of the piecewise defined function that expresses the final macro-averaged F-measure as a function of train-test ratio for standard "non-grouping" classification "control". The

learning rate (slope) from the linear equation 1 (for the cases of training set size less than test set size) to 2 (training set size greater than test set size) was significantly reduced by a factor of 50 for small categories, 47 for medium categories, and 99 for large categories. The overall results supported that the performance was getting close to its asymptote (i.e., maximum level of performance) when increasing the training dataset that outnumbered test dataset. Large categories were much closer to its asymptotic performance than smaller categories for the standard classification scenario.

Table 8.8: Slope and Intercept of Linear Equations Relating Train-Test Ratio and Macro-averaged F-measure

	S		**M**		**L**	
Slope	**1**	2	**1**	2	**1**	2
MNB	13.71%	0.47%	10.59%	0.28%	1.89%	0.01%
SVM	15.97%	0.19%	5.47%	0.09%	3.16%	0.04%
LR	19.30%	0.31%	5.38%	0.08%	2.37%	0.03%
Average	16.33%	0.32%	7.15%	0.15%	2.47%	0.03%

	S		**M**		**L**	
Intercept	**1**	2	**1**	2	**1**	2
MNB	14.21%	27.45%	55.28%	65.59%	77.36%	79.25%
SVM	52.85%	68.63%	71.69%	77.07%	79.82%	82.94%
LR	47.27%	66.26%	72.66%	77.96%	81.12%	83.46%
Average	38.11%	54.11%	66.54%	73.54%	79.43%	81.88%

Given the piecewise defined function that relates the performance and train-test ratio for standard classification of each combined setting of category size and classifier, the expected train-test ratio r required to achieve the performance level F can be calculated accordingly.

$$r(F) = \begin{cases} (F - intercept_1)/slope_1; & F_1 \leq F < F_2 \\ (F - intercept_2)/slope_2; & F_2 \leq F < F_3 \end{cases} \tag{8.6}$$

where F_1, F_2, and F_3 are the performance levels of standard classification performance of the same classifier and category size at the train-test ratio of $\frac{1}{9}$, 1, and 9.

To make the improvement of the proposed Grouping Methods more relatable, the percent-increase can be converted into the additional training data that is required to achieve the same level of performance. Given the same setting of classifier and category size, the expected train-test ratio for the macro-averaged F-measure achieved by a specific Grouping Method is first calculated. By comparing the actual train-test ratio r to the expected train-test ratio r', the positive grouping effect can be represented by the additional training required to achieve that performance level quantified in terms of its proportion to the actual training dataset size. Positive Grouping Effect in terms of training data saved is calculated by:

$$\frac{(r' - r)}{r}$$

where r is the actual train-test ratio and r' is the expected train-test ratio.

As noted earlier, the automated Type M+S Grouping Method plus two add-on methods: "S2" for two-word sequence tagging and "PA" for reviewed tagging of PA-concept ("MSS2PA") was the overall best-performing grouping method for SVM and MNB. Pairing this method with the "NB" add-on for supplying Naive Bayes Weighted feature values ("MSS2PA_NB") was superior for LR.

The proposed grouping method tended to have greater effect on small categories. When the training dataset is limited compared to the test set (train-test ratio of 1:9), the proposed Word Grouping Method was able to improve the classification performance of MNB by 5.5%, SVM by 4%, and LR by 11.2%, which are

comparable to increasing the size of the labeled training set by a factor of 3.6 for MNB, 2.3 for SVM, and 5.2 for LR.

The results also showed that when the size of the training set was comparable to the test set (train-test ratio of 1:1), huge, unrealistic amounts of additional training data were required to reach the same performance level that was achieved by the proposed grouping method. As Table 8.9 suggests, the additional training set may have to be at least 12 times larger than the current to achieve that level of improvement, which unfortunately is not feasible in practice.

By visually examining the approaching asymptotic performance of standard classification in Figure 8.10, one may realize that simply increasing training data may still never reach the same performance level as the proposed word grouping.

Table 8.9: Grouping Effect on Improving Small Categories in Terms of Saved Training Data Proportional to Current Training Dataset Size

Classifier	**1:9**	**1:1**	**9:1**	**Overall**	**Grouping Method**
MNB	3.6	14.4	1.8	6.6	MSS2PA
SVM	2.3	12.8	0.6	5.2	MSS2PA
LR	5.2	19.7	1.3	8.7	MSS2PA_NB
Average	3.7	15.7	1.2	6.9	

To sum up, the proposed Grouping Models were effective in improving the overall classification performance for three classifiers in three train-test ratio scenarios. Some proposed methods such as "MSS2PA" and "MSS2PA_NB" were statistically more effective than the classic word grouping methods of stemming and lemmatization. The proposed Grouping Models were more effective in the case of limited training samples (small categories and small train-test ratio). The results proved the effectiveness of utilizing and grouping rare and unseen features that stemming and lemmatization failed to address (i.e., misspellings, domain-specific

words, and same-hypernym words with dissimilar spelling) and the incorporated add-on methods of two-word sequence tagging, reviewed tagging, and applying Naive Bayes-weighted LR.

Last but not least, Hypotheses 6 and 7 were tested accordingly. Given that the proposed Grouping Models of "MSS2PA" and "MSS2PA_NB" outperformed stemming and lemmatization in terms of the overall classification performance with statistical significance, Null Hypothesis 6 was falsified and Alternative Hypothesis 6 was accepted. Hypothesis 7 concerns the effectiveness of the add-on methods: considering sequences of words, involving manual review, and applying Naive Bayes-weighted SVM and LR. The three add-on methods were found effective in improving the Type M+S Grouping Method, except for applying Naive Bayes-weighted SVM. Thus, Null Hypothesis 7 was falsified and Alternative Hypothesis 7 was accepted conditionally.

8.3 Conclusion

While being amazed by how well simple machine learning and statistical models work in classifying textual data, researchers have gradually realized that these automated statistical models can only get to a certain level of accuracy (roughly 70% from a previous study of Nanda and Lehto on another injury dataset) because of their inherent limitation to address data rarity and imbalance (limited labeled examples in small categories of an imbalanced dataset or in small training set compared to test set). The high dimensionality and sparsity of Vector Space Model (VSM) of textual data have prevented machine learning models from full learning of meaningful but rare patterns. In response to these two primary challenges in statistical text analysis, researchers in relevant areas have focused primarily on improving algorithms, while few studies have been done to examine and improve these statistical methods from a linguistic perspective.

Due to the lack of discriminatory power (Luhn, 1958) and significant contribution to the size of vocabulary and total word occurrences in a corpus (Zipf, 1949), high- and low- frequency words are often removed in statistical text analysis. However, some studies have demonstrated the benefit of keeping these extreme-frequency words in statistical text analysis. To answer the fundamental question of whether to remove or keep extreme-frequency features in text classification of injury surveillance data, this study systematically examined the removal effect on the classification performance of injury narratives for three classifiers (Multinomial Naive Bayes "MNB", Support Vector Machine "SVM", Logistic Regression "LR") in three train-test ratio scenarios (1:9, 1:1, 9:1).

Since some high-frequency words are essential to classification, the high-frequency words in examination were limited to the five types of stopwords (grammatical articles, pronouns, auxiliary verbs, prepositions, others). This study showed that removing stopwords had a positive impact on overall classification performance. Although their removal had a positive effect on overall classification performance (+0.6%), preposition-stopwords should not be removed because their absence greatly negatively impacted the classification of some categories (e.g. PEDESTRIAN; -3.6%). Consistent with previous research (Riloff, 1995), this study also supported that, despite their lack of semantic meaning, some stopwords are indeed necessary for building complex phrases and providing specific concepts, which may not be achievable by the pairing word alone.

The importance of low-frequency words (LFWs) in classifying injury data was also explored in this study. Removing LFWs significantly improved overall classification performance for MNB. As MNB is one of the earliest developed classifiers, its significant positive effect may justify the advocacy of removing LFWs in early research. However, the more recently developed classifiers, SVM and LR, were insignificantly impacted by the LFW removal (less than 1%). Interestingly, removing LFWs had a positive impact on small categories for LR. As LR is prone to overfitting with limited training data, this finding can be explained as an artifact of over-

fitting. Removing LFWs may address the overfitting problem of LR, allowing the model to learn from better-represented Vector Space Models (VSMs) by distributing weights more evenly to other frequent, discriminative features, rather than relying on rare features that may never occur in prediction data. Thus, the overall results seem to support the widely-held belief that LFWs from injury narratives can be removed with no-to-minimum negative impact. In addition to the benefit of the reduced dimension of feature space (vocabulary) and improved computational efficiency, removing LFWs showed the significant improvement on overall classification performance for MNB, insignificant impact for SVM, and potential benefit in alleviating the overfitting problem for LR.

In addition to examining extreme-frequency words in the text classification of injury narratives, the second objective of this study is to utilize LFWs for potential classification improvement. LFWs need careful handling because they contribute to the majority of vocabulary (half may be composed of words that occurred only once) and removing them without grounding is considered ad hoc approach (Yang & Pedersen, 1997). Words have different forms for grammatical reason, but they carry similar or related meaning with their root or base form. LFWs are considered important if they share the same hypernym (similar concept) with some discriminative words by being their morphological variants, misspellings, or even synonyms with dissimilar spelling. An ideal word normalization or grouping method should group same-hypernym words either with similar or dissimilar spelling. United representative forms can improve statistical robustness of discriminatory features and thus create a smaller and denser but more representative VSM for training a classifier to make accurate predictions. The classic word grouping methods of stemming or lemmatization are often implemented to group same-hypernym words with similar spelling by removing the ending of words; however, they have limitations to address misspellings, domain-specific words, or synonyms with dissimilar spelling.

In this study, the proposed Type M+S Grouping Method comprises two parts: "Type-M Morphological Mapping" for grouping words with similar spelling and "Type-S Semantic Tagging" for grouping words with dissimilar spelling. To minimize manual effort, several existing statistical methods were utilized in a mix-and-match fashion for nonconventional purposes, identifying (1) predictive categories using the coefficient matrix of a linear classifier; (2) same-hypernym words with given discriminatory seedwords using the statistical semantic measure of Word2Vec; and (3) discriminatory features as seedwords for comparing with other words using feature selection methods. The proposed method utilized rare and unseen words through linguistic grouping by their morphological and semantic similarity measured statistically. Optimized by the add-on methods of two-word sequence tagging, reviewed tagging, and Naive Bayes-weighted input features, the proposed Type M+S Grouping Method statistically significantly outperformed the classic grouping methods of stemming and lemmatization. This demonstrates the effectiveness of (1) considering unseen words, which are often neglected or removed in practice; (2) statistically considering linguistic features (morphology, semantics, two-word sequences); and (3) incorporating human semantic knowledge in the form of manual review.

Serving as a means of decision support, the proposed Word Grouping method is a promising approach for incorporating expert knowledge that improves machine learning for classifying injury narratives with reduced manual effort. The results also suggest that simply increasing the size of a training set would not result in the level of performance that the proposed method can achieve because of the inherent limitations of linear classifiers to acquire fundamental concepts and classification rules from the narrative that human experts know by definitions of injuries.

8.4 Future Work

This study aimed to improve statistical machine learning methods from a linguistic perspective while utilizing various statistical techniques in natural language processing to minimize the manual effort of supplying annotated semantics. This study confirmed the potential of Word2Vec as a measure of statistical semantics in capturing the meaning of human natural language while also demonstrating its limitations and how human review can improve the quality of the results generated by Word2Vec in an unsupervised way.

Strategic utilization of statistical methods can serve as a means of cross validation to reduce manual review effort. In safety research, many studies that utilized machine learning approaches to classify injury narratives have showed the potential for prioritizing manual review effort by examining the predictive probability strength of classifiers as a confidence metric or the agreement between multiple algorithms (Marucci-Wellman, Corns, & Lehto, 2017). Specifically, these filtering strategies filter out and review the narratives that have low predictive strength or conflicting predictions by different classifiers.

Similar strategies can be applied to Word2Vec. Word2Vec is an unsupervised learning approach, which does not require any labeled data. However, if labeled data are available, strategically training several Word2Vec models on different samples of corpus may provide a good basis for cross-validating the derived results and improve the quality while reducing manual review effort. For example, in Section 7.2.2, a Word2Vec model was trained on the entire QISU dataset for the high recall of identifying words that were similar to any of the selected seedwords: drug, chemical, alcohol, food, or plant. Since these seedwords were common types of injury agents and share similar context in injury narratives of the QISU dataset, words similar to any of these seedwords could also be similar to other seedwords in terms of Word2Vec. For example, in the QISU dataset, the word most similar to "drug" in Word2Vec is not a drug, but instead an alcohol:

"absynth," which is the misspelling of "absinthe" (i.e., a distilled, highly alcoholic beverage). The exploratory studies in Section 7.2.1 provided more examples and evidence that Word2Vec tended to favor rare features, including meaningful low-frequency synonyms (a benefit) and random co-occurrences (a drawback) simultaneously. Realizing that the quality of Word2Vec could be further improved by manual review, the proposed method allowed for the incorporation of human semantic knowledge in a form of manual review. The manual review effort may be reduced by strategically training different Word2Vec models on different topic-specific samples of corpus, for example, narratives that belong to a specific external cause category of PA_DRUG, PA_ALCOHOL, PA_CHEMICAL, PA_FOOD, and PA_PLANT. The Word2Vec model that is trained on a topic-specific dataset should have a higher precision but possibly lower recall of identifying same-hypernym (similar-concept) words and thus may compensate the general Word2Vec model that is trained on the entire dataset for its low precision by cross-validating the results and excluding the ones that both agree from a to-review list.

In addition to supplying semantic knowledge in a form of manual review, human experts can contribute to the automated machine learning process with domain knowledge to develop classification rules based on the understanding for the corpus being analyzed. An injury corpus is composed of a variety of simple or complex concepts that involve single-token words, (contiguous or non-contiguous) word sequences, and word combinations. The classic word grouping methods of stemming and lemmatization are limited to cope with the same-hypernym words with similar spellings. However, most discriminative concepts in an injury corpus are composed of synonyms with entirely different spellings, which are partially addressed by the utilization of statistical semantics in this study, and complex, multi-word expressions, which are often difficult for statistical methods to acquire without human intervention. The understanding of the fundamental concepts in the corpus being analyzed is essential to identify discriminatory concepts and represent training instances accordingly for training a reliable classifier. The

literature has acknowledged the importance of human-machine collaboration to integrate human semantics or domain knowledge in a form of ontology into statistical models for improving automated text analysis (Baharudin, Lee, & Khan, 2010; Somprasertsri & Lalitrojwong, 2010; Yu & Dang, 2012; Zhao & Li, 2009). An ontology is defined as "effectively formal and explicit specifications in the form of concepts and relations of shared conceptualizations" (Gruber, 1993; Wong, Liu, & Bennamoun, 2012). Simply put, an ontology provides a controlled, hierarchical vocabulary of concepts, each with explicitly defined and machine-readable semantics. The purpose of ontologies is to provide machines with structured knowledge about specific domains so that they can be trained properly on a well-represented Vector Space Model. Wong et al. (2012) suggested that an ontology has four elements in hierarchy: Terms, Concepts, Relations, and Rules. Each ontological element is a prerequisite for obtaining the element of the next layer. This study was able to reach the second level of Concepts by utilizing statistical methods to develop synonym lists for discriminatory concepts. Future work is suggested to explore the feasibility of human-machine collaboration in acquiring the higher-level ontological elements of Relations and Rules. Simple linguistic rules are expected to greatly improve the recall of certain small unique categories. For example, a study indicated that an additional 44% of cases from the external cause category of "electrocution" were identified using expert-specified keywords (Marucci-Wellman et al., 2017). The initial exploration in the QISU dataset also showed that the recall was improved for the FIREARM category from 18% to 68% by using simple keywords (e.g. $rifle$, $bullet$, $shot\_gun$, $spear\_gun$) and from 74% to 92% for DROWNING using less than ten simple linguistic rules (such as the presence of "child" with "under water" or "cpa" with "difficult breathing.")

Therefore, further study on this path is encouraged, namely the development of a collection of expert rules for identifying essential concepts related to the safety and injury domain by utilizing expert knowledge paired with statistical- and linguistics- based techniques. This collection of rules will serve as a con-

trolled vocabulary or ontology that is considered a lightweight version of ontological semantics proposed by Nirenburg and Raskin (2004) with a main focus of identifying and normalizing discriminatory concepts from injury narratives. This sharable, domain-specific ontology should greatly improve the quality of statistical text analysis in the safety and injury area by addressing the limitations of statistical models in disambiguating word senses and acquiring concepts that are low in frequency, high in variety, and complex in expression.

LIST OF REFERENCES

LIST OF REFERENCES

Addiction Prevention Center. (2008). *Psychoactive Medication. Drugs: Know the Facts, Cut Your Risk.* Retrieved from http://www.toxquebec.com/livre_drogues/en/index_medicaments.html

Al-Tahrawi, M. M. (2013). The role of rare terms in enhancing the performance of polynomial networks based text categorization. *Journal of Intelligent Learning Systems and Applications*, *5*(2), 84–89.

Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, *1*(1), 4–20.

Baofu, P. (2011). *The Future of Post-Human Chemistry: A Preface to a New Theory of Substances and their Changes*. Cambridge Scholars Publishing, UK.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. , 238–247.

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Berry, M. W., Drmac, Z., & Jessup, E. R. (1999). Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, *41*(2), 335–362.

Bertke, S. J., Meyers, A. R., Wurzelbacher, S. J., Measure, A., Lampl, M. P., & Robins, D. (2016). Comparison of methods for auto-coding causation of injury narratives. , *88*, 117–123.

Bird, S., Loper, E., & Klein, E. (2009). *Natural language processing with python*. OReilly Media Inc.

Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. In P. Buitelaar, P. Cimiano, & B. Magnini (Eds.), *Ontology learning from text: Methods, evaluation and applications* (pp. 1–10). Amsterdam, The Netherlands: IOS Press.

Chennuru, S., Chen, P.-W., Zhu, J., & Zhang, J. Y. (2012). Mobile Lifelogger Recording, Indexing, and Understanding a Mobile User's Life. In (pp. 263–281). Springer Berlin Heidelberg.

Choe, P., Lehto, M., Shin, G.-C., & Choi, K.-Y. (2013). Semiautomated Identification and Classification of Customer Complaints. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *23*(2), 149–162.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

CTI Reviews. (2016). *Drugs, Society, and Human Behavior.* Retrieved from http://www.nhs.uk/chq/Pages/859.aspx?CategoryID=73

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407.

Delavenay, E. (1960). *An introduction to machine translation*. New York, NY: An Introduction to Machine Translation.

Dirkmaat, D. (2013). *e-Study Guide for: A companion to forensic anthropology*. Wiley-Blackwell.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61-74.

Ertekin, S., Giles, C. L., Storage, H. I., & Miscellaneous, R. (2007). Active Learning for Class Imbalance Problem. In *Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval.* (pp. 823–824).

Essen, U., & Steinbiss, V. (1992). Cooccurrence Smoothing for Stochastic Language Modeling. In *Proceedings of the ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 161–164). San Francisco, CA: IEEE.

Farlex Partner Medical Dictionary. (2012). *Recreational drug use.* Retrieved from http://medical-dictionary.thefreedictionary.com/Recreational+drug+use

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford, UK: Philological Society.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

GreenFacts. (2016). *Psychoactive Drugs Tobacco, Alcohol, and Illicit Substances.* Retrieved from http://www.greenfacts.org/en/psychoactive-drugs/1-2/3-drug-addiction-brain.htm\#3

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Han, B., Cook, P., & Baldwin, T. (2012). Automatically Constructing a Normalisation Dictionary for Microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, (pp. 421–432). Jeju Island, Korea: Association for Computational Linguistics.

Han, L., Finin, T., McNamee, P., Joshi, A., & Yesha, Y. (2013). Improving word similarity by augmenting PMI with estimates of word polysemy. In *Ieee transactions on knowledge and data engineering* (Vol. 25, pp. 1307–1322).

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Holte, R. C. (1989). Concept learning and the problem with small disjuncts. In *Proceedings of 11th international joint conference on artificial intelligence* (pp. 813–818). Detroit, MI: Morgan Kaufmann.

Huang, H. Y., Nanda, G., Lehto, M., & Vallmuur, K. (2016). Extracting name of injury agent using semantic data mining method. In *Proceedings of the 7th international conference on applied human factors and ergonomics*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lehto, M., Marucci-Wellman, H., & Corns, H. (2009). Bayesian methods: a useful tool for classifying injury narratives into cause groups. *Injury Prevention*, *15*(4), 259-265.

Lehto, M., & Sorock, G. S. (1996). Machine learning of motor vehicle accident categories from narrative data. *Methods of information in medicine*, *35*(4-5), 309–16.

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th european conference on machine learning* (pp. 4–15). Chemnitz, Germany: Springer Berlin Heidelberg.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics* (pp. 768–774). Montreal, Quebec, Canada: Association for Computational Linguistics.

Liu, B., Hsu, W., & Ma, Y. (1999). Mining association rules with multiple minimum supports. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 337–341). San Diego, CA: ACM Press.

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class imbalance problem in data mining: Review. *International Journal of Computer Science and Network*, *2*(1), 83–87.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, *2*(2), 159–165.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Manning, C. D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In *International conference on intelligent text processing and computational linguistics* (pp. 171–189). Springer Berlin Heidelberg.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

Marucci-Wellman, H., Corns, H., & Lehto, M. (2017). Classifying injury narratives of large administrative databases for surveillance A practical approach combining machine learning ensembles and human review. *Accident analysis and prevention*, *98*, 359 - 371.

Marucci-Wellman, H., Lehto, M., & Corns, H. (2011). A combined Fuzzy and Naive Bayesian strategy can be used to assign event codes to injury narratives. *Injury Prevention*, *17*(6), 407-414.

Marucci-Wellman, H., Lehto, M., & Corns, H. L. (2015). A practical tool for public health surveillance: Semi-automated coding of short injury narratives from large administrative databases using Nave Bayes algorithms. *Accident Analysis and Prevention*, *84*, 165-176.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *Proceeding of the AAAI-98 workshop on learning for text categorization* (pp. 41–48). CA: AAAI Press.

McKenzie, K., Scott, D. A., Campbell, M. A., & McClure, R. J. (2010). The use of narrative text for injury surveillance research: a systematic review. *Accident Analysis and Prevention*, *42*(2), 354-363.

MD Medical Reference. (2016). *What Is Sinusitis?* Retrieved from http://www.webmd.com/allergies/guide/what-is-sinusitis?page=2

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Montgomery, D. C. (2009). *Design and analysis of experiments*. Wiley.

Nanda, G., Grattan, K. M., Chu, M. T., Davis, L. K., & Lehto, M. (2016). Bayesian decision support for coding occupational injury data. *Journal of Safety Research*, *57*, 71-82.

National Institute on Drug Abuse. (1969). *Cocaine.* Retrieved from https://www.drugabuse.gov/publications/drugfacts/cocaine

News Medical. (2007). Mixing certain allergy medications with other medicines can have hazardous effects on your health. Retrieved from http://www.news-medical.net/news/2007/04/24/24084.aspx

NHS Choice. (2015). *Can I take paracetamol if I'm on antibiotics?* Retrieved from http://www.nhs.uk/chq/Pages/859.aspx?CategoryID=73

Nirenburg, S., & Raskin, V. (2004). *Ontological semantics*. The MIT Press.

Northern Territory Government Health Services. (n.d.). *An overview of alcohol and other drug issues.* Retrieved from http://www.nt.gov.au/health/healthdev/health_promotion/bushbook/volume2/chap1/sect1.htm.

Ogbru, O. (2015). *NoCetirizine, Zyrtec, Zyrtec Allergy, Zyrtec Hives.* Retrieved from http://www.medicinenet.com/cetirizine/article.htm

Pagallo, G., & Haussler, D. (1990). Boolean feature discovery in empirical learning. *Machine Learning*, *5*, 71–99.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the international conference on language resources and evaluation, LREC* (Vol. 10, pp. 1320–1326).

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (Vol. 41, pp. 613–619). ACM.

Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL student research workshop* (pp. 13–18). Ann Arbor, Michigan: Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825-2830.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137.

Price, L., & Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology*, *56*(8), 883–888.

Psychology Encyclopedia. (2016). *Psychoactive Drugs - Overview and use, Side effects, Precautions.* Retrieved from http://psychology.jrank.org/pages/509/Psychoactive-Drugs.html

PubMed Health. (2015). *Drug names and classes.* Retrieved from https://www.ncbi.nlm.nih.gov/pubmedhealth/drug-names-and-classes/

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - (pp. 519–526). Morristown, NJ, USA: Association for Computational Linguistics.

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th international conference on Computational linguistics* (Vol. 2, pp. 821–827). Taipei, Taiwan: Association for Computational Linguistics.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth machine translation summit* (pp. 315–322).

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. In *Proceedings of the 4th international conference on language resources and evaluation (LREC 2004)* (pp. 395–398).

Rehurek, R., & Sojka, P. (2010). *Gensim: topic modeling for humans.* Retrieved from https://radimrehurek.com/gensim/models/word2vec.html

Rei, L. (2015). *Multiclass Naive Bayes SVM (NB-SVM).* Retrieved from https://github.com/lrei/nbsvm

Riloff, E. (1995). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 130–136).

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing & Management*, *28*(3), 317-332.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words*. Stockholm Universit.

Saif, H., Fernandez, M., He, Y., & Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In *The 9th international conference on language resources and evaluation.* Reykjavik, Iceland.

Salton, G. (1971). *The SMART retrieval system - experiments in automatic document processing*. Upper Saddle River, NJ: Prentice-Hall.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513-523.

Salton, G., Wong, A., & Yang, C.-S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, *18*(11), 613–620.

Saussure, F. d. (1916). *Cours de linguistique gé- nérale (Course in general Linguistics)* (C. Bally & A. Sechehaye, Eds.). Paris: McGraw-Hill.

Schnhofen, P., & Benczr, A. A. (2006). Exploiting extremely rare features in text categorization. In *Proceeding of 17th european conference on machine learning* (pp. 759–766). Berlin, Germany: Springer Berlin Heidelberg.

Schtze, H. (1995). Distributional part-of-speech tagging. In *Proceedings of the 7th conference of the european chapter of the association for computational linguistics (EACL 1995)* (pp. 141–148).

Schtze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Simeon, M., & Hilderman, R. (2008). Categorical Proportional Difference: A Feature Selection Method for Text Categorization. In *Proceeding of the seventh australasian data mining conference (ausdm 2008)* (pp. 201–208). Glenelg, South Australia.

Skwarecki, B. (2015). *Advil Beats Panadol for Pain Relief.* Retrieved from `http://www.lifehacker.com.au/2015/08/advil-beats-panadol-for-pain-relief/`

Somprasertsri, G., & Lalitrojwong, P. (2010). Mining feature-opinion in online customer reviews for opinion summarization. *Journal of Universal Computer Science*, *16*(6), 938–955.

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21.

Sparck Jones, K. (1973). Index Term Weighting. *Information Storage and Retrieval*, *9*(11), 619–633.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology* (Vol. 1, pp. 173–180). Morristown, NJ, USA: Association for Computational Linguistics.

Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416.

Turney, P., Littman, M., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the international conference on recent advances in natural language processing* (pp. 482–489).

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(2010), 141–188.

U.S. Food and Drug Administration. (2016). *Nonsteroidal Anti-inflammatory Drugs (NSAIDs).* Retrieved from http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm103420.htm

Vallmuur, K. (2015). Machine learning approaches to analysing textual injury surveillance data: a systematic review. *Accident Analysis and Prevention*, *79*, 41–49.

van Rijsbergen, & J., C. (1979). *Information Retrieval (2nd ed.)*. London, UK: Butterworths.

Versley, Y. (2008). Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In *Proceedings of the ESSLLI workshop on distributional lexical semantics* (pp. 55–62).

Weaver, W. (1955). Translation. In W. N. Locke & A. D. Booth (Eds.), *Machine translation of languages* (Vol. 14, pp. 15–23). Cambridge, MA: MIT Press.

Weiss, G. M. (1995). Learning with rare cases and small disjuncts. In *Proceedings of 12th international conference on machine learning* (pp. 558–565). Tahoe City, California: Morgan Kaufmann.

Weiss, G. M. (2004). Mining with rarity: A unifying framework. *ACM SIGKDD Explorations Newsletter*, *6*(1), 7–19.

Weiss, G. M. (2005). Mining with rare cases. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook: A complete guide for practitioners and researchers* (pp. 765–776).

Weiss, G. M., & Hirsh, H. (2000). A quantitative study of small disjuncts. In *Proceeding of 17th national conference on artificial intelligence* (pp. 665–670). Austin, TX: AAAI Press.

Wellman, H., Lehto, M., Sorock, G. S., & Smith, G. S. (2004). Computerized coding of injury narrative data from the National Health Interview Survey. *Accident; Analysis and Prevention*, *36*(2), 165–171.

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text. *ACM Computing Surveys*, *44*(4), 1–36.

World Health Organization. (2004). *Neuroscience of psychoactive substance use and dependence.* Retrieved from http://www.who.int/substance_abuse/publications/en/Neuroscience_E.pdf

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd international ACM SIGIR conference on research and development in information retrieval* (pp. 42–49). Berkeley, CA: ACM Press.

Yang, Y., & Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning* (pp. 412–420).

Yu, Y., & Dang, J. (2012). Semantic mining on customer survey. In *Proceedings of the 8th international conference on semantic systems* (pp. 72–79). Graz, Austria: ACM.

Zhao, L., & Li, C. (2009). Ontology based opinion mining for movie reviews. In D. Karagiannis & Z. Jin (Eds.), *Knowledge science, engineering and management* (pp. 204–214).

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Boston, MA: Addison-Wesley.

APPENDICES

A. SUPPLEMENTARY TABLES FOR STOPWORD REMOVAL EXPERIMENTS

Table A.1: ANOVA Table (F-measure): Five Types of Stopwords vs. Control

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	42	88.0482887	2.0963878	239.49	<.0001
Error	1577	13.8043706	0.0087536		
Corrected Total	1619	101.8526594			

Source	DF	Type I / III SS	Mean Square	F Value	Pr >F
Category	29	69.9495149	2.41205224	275.55	<.0001
Classifier	2	14.74421229	7.37210614	842.18	<.0001
StopwordTypeRemoval	5	0.00660723	0.00132145	0.15	0.9798
Train-Test_Ratio	2	3.26916294	1.63458147	186.73	<.0001
Classifier*Ratio	4	0.07879137	0.01969784	2.25	0.0616

Table A.2: ANOVA Table (Impact): Five Types of Stopwords

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	46	0.062311	0.001355	20.23	<.0001
Error	1303	0.087256	0.000067		
Corrected Total	1349	0.149567			

Source	DF	Type III SS	Mean Square	F Value	Pr >F
CategorySize	2	0.00541	0.00271	40.4	<.0001
Classifier	2	0.00243	0.00121	18.12	<.0001
RemovedStopwordType	**4**	**0.00149**	**0.00037**	**5.56**	**0.0002**
Train-Test_Ratio	2	0.00042	0.00021	3.16	0.0427
Classifier*CategorySize	4	0.00373	0.00093	13.93	<.0001
RemovedStopwordType*CategorySize	8	0.00917	0.00115	17.12	<.0001
Classifier*RemovedStopwordType	8	0.00522	0.00065	9.74	<.0001
RemovedStopwordType*Classifier*CategorySize	16	0.01137	0.00071	10.61	<.0001

Table A.3: Post Hoc Tests (Tukey and LSD) for Removed Stopword Types

Tukey Grouping		Mean	N	RemovedStopwordType	LSD Grouping	
	A	0.0059	270	preposition		A
	B	0.00246	270	other		B
C	B	0.00116	270	aux. verb	C	B
C	B	0.0007	270	pronoun	C	
C		0.00026	270	article	C	

B. SUPPLEMENTARY TABLES FOR LFW REMOVAL EXPERIMENTS

Table B.1: ANOVA Table (F-measure): Removing DF1-9 LFW vs. Control

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	118	125.2600147	1.0615255	200.41	<.0001
Error	2581	13.6709034	0.0052967		
Corrected Total	2699	138.930918			

Source	DF	Type I / III SS	Mean Square	F Value	Pr >F
Category	29	111.1486015	3.8327104	723.6	<.0001
Classifier	2	7.8202771	3.9101385	738.22	<.0001
DFC	9	0.2760213	0.030669	5.79	<.0001
Train-Test_Ratio	2	5.2584328	2.6292164	496.38	<.0001
Classifier*DFC*Ratio	76	0.756682	0.0099563	1.88	<.0001

Table B.2: ANOVA Table (F-measure): Removing DF1-9 LFW vs. Control for MNB

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	58	58.10787186	1.00186	279.99	<.0001
Error	841	3.00924966	0.003578		
Corrected Total	899	61.11712152			

Source	DF	Type I / III SS	Mean Square	F Value	Pr >F
Category	29	55.22115193	1.904178	532.16	<.0001
DFC	9	0.98015856	0.108907	30.44	<.0001
Train-Test_Ratio	2	1.88642748	0.943214	263.6	<.0001

Table B.3: ANOVA Table (F-measure): Removing DF1-9 LFW vs. Control for SVM

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	58	30.37266446	0.52366663	330.97	<.0001
Error	841	1.33065264	0.00158223		
Corrected Total	899	31.7033171			

Source	DF	Type I / III SS	Mean Square	F Value	Pr >F
Category	29	28.79000758	0.99275888	627.44	<.0001
DFC	9	0.00600375	0.00066708	0.42	0.924
Train-Test_Ratio	2	1.57131668	0.78565834	496.55	<.0001

Table B.4: ANOVA Table (F-measure): Removing DF1-9 LFW vs. Control for LR

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	58	36.56632967	0.63045396	307.57	<.0001
Error	841	1.72387268	0.00204979		
Corrected Total	899	38.29020235			

Source	DF	Type I / III SS	Mean Square	F Value	Pr >F
Category	29	34.74457039	1.19808863	584.49	<.0001
DFC	9	0.00109204	0.00012134	0.06	1
Train-Test_Ratio	2	1.81981513	0.90990757	443.9	<.0001

Table B.5: Post Hoc Tests (Tukey and LSD): Removing DF1-9 LFW vs. Control for MNB

Means with the same letter are not significantly different.						
Tukey Grouping		**F-measure Mean**	**N**	**DFC**	**LSD Grouping**	
	A	0.617994	90	10		A
	A	0.61673	90	9		A
	A	0.615422	90	8		A
	A	0.613918	90	7		A
	A	0.611671	90	6		A
	A	0.607434	90	5	B	A
	A	0.6009	90	4	B	A
B	A	0.590081	90	3	B	
B		0.571078	90	2		C
	C	0.505273	90	1		D

Table B.6: Overall Classification Performance at DFC Levels 1 to 10

Macro-averaged F-measure		**Document Frequency Cut-off (DFC)**										
Train-Test Ratio	Classifier	1	2	3	4	5	6	7	8	9	10	Overall
1:9	MNB	42.93%	49.77%	52.28%	53.64%	54.56%	55.18%	55.6%	55.85%	55.97%	56.08%	53.19%
	SVM	66.34%	66.02%	65.74%	65.60%	65.42%	65.24%	65.03%	64.83%	64.67%	64.44%	65.33%
	LR	65.06%	65.08%	65.06%	65.00%	64.92%	64.85%	64.76%	64.63%	64.49%	64.36%	64.82%
1:1	MNB	53.16%	59.37%	61.26%	62.05%	62.54%	62.87%	63.03%	63.15%	63.36%	63.46%	61.43%
	SVM	72.40%	72.13%	72.18%	72.17%	72.22%	72.07%	71.99%	71.85%	71.76%	71.73%	72.05%
	LR	72.42%	72.46%	72.46%	72.44%	72.47%	72.52%	72.39%	72.37%	72.35%	72.22%	72.41%
9:1	MNB	55.49%	61.89%	63.44%	64.55%	65.06%	65.43%	65.56%	65.69%	65.81%	65.94%	63.89%
	SVM	75.50%	75.39%	75.31%	75.29%	75.40%	75.27%	75.41%	75.40%	75.44%	75.40%	75.38%
	LR	75.51%	75.54%	75.57%	75.59%	75.54%	75.53%	75.54%	75.52%	75.49%	75.49%	75.53%

Table B.7: Overall Effect of Removing LFWs on Classification Performance at DFC Levels 2 to 10

Impact		**Document Frequency Cut-off (DFC)**									
Train-Test Ratio	**Classifier**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Overall**
1:9	MNB	6.84%	9.35%	10.71%	11.63%	12.25%	12.67%	12.92%	13.04%	13.15%	11.40%
	SVM	-0.32%	-0.60%	-0.74%	-0.92%	-1.10%	-1.31%	-1.51%	-1.67%	-1.90%	-1.12%
	LR	0.02%	0.00%	-0.06%	-0.14%	-0.21%	-0.30%	-0.43%	-0.57%	-0.70%	-0.27%
1:1	MNB	6.21%	8.10%	8.89%	9.38%	9.71%	9.87%	9.99%	10.20%	10.30%	9.18%
	SVM	-0.27%	-0.22%	-0.23%	-0.18%	-0.33%	-0.41%	-0.55%	-0.64%	-0.67%	-0.39%
	LR	0.04%	0.04%	0.02%	0.05%	0.10%	-0.03%	-0.05%	-0.07%	-0.20%	-0.01%
9:1	MNB	6.40%	7.95%	9.06%	9.57%	9.94%	10.07%	10.20%	10.32%	10.45%	9.33%
	SVM	-0.11%	-0.19%	-0.21%	-0.10%	-0.23%	-0.09%	-0.10%	-0.06%	-0.10%	-0.13%
	LR	0.03%	0.06%	0.08%	0.03%	0.02%	0.03%	0.01%	-0.02%	-0.02%	0.02%

Table B.8: Overall Classification Performance for Different Category Sizes at DFC Levels 1 to 10

(a) MNB

% F-measure		DF Cutoff										
Train-Test Ratio	**category Size**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**overall**
1:9	**L**	77.57	78.05	78.04	78.02	77.99	77.96	77.94	77.92	77.9	77.88	77.93
	M	56.45	62.44	64.35	65.34	65.95	66.34	66.59	66.81	66.93	67.03	64.82
	S	15.73	25.04	28.95	31.12	32.7	33.79	34.55	34.91	35.03	35.2	30.7
1:1	**L**	78.77	78.97	78.94	78.91	78.89	78.87	78.84	78.83	78.81	78.79	78.86
	M	65.45	69.76	70.79	71.04	71.15	71.18	71.24	71.22	71.27	71.21	70.43
	S	29.53	39.75	43.33	45.09	46.27	47.12	47.48	47.82	48.33	48.7	44.34
9:1	**L**	79.29	79.24	79.16	79.11	79.07	79.05	79.03	79.01	78.99	78.98	79.09
	M	68.11	71.91	72.37	72.56	72.58	72.61	72.63	72.59	72.58	72.54	72.05
	S	31.65	43.24	46.78	49.53	50.89	51.86	52.19	52.6	52.95	53.37	48.51

(b) SVM

% F-measure		DF Cutoff										
Train-Test Ratio	**category Size**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**overall**
1:9	**L**	80.17	80.21	80.28	80.35	80.42	80.51	80.59	80.65	80.71	80.75	80.46
	M	72.3	72.1	71.94	71.84	71.71	71.64	71.59	71.52	71.44	71.38	71.75
	S	54.62	54.05	53.51	53.26	52.97	52.59	52.07	51.63	51.3	50.76	52.68
1:1	**L**	82.06	82.1	82.16	82.21	82.26	82.29	82.33	82.34	82.36	82.38	82.25
	M	75.49	75.41	75.38	75.42	75.45	75.42	75.46	75.42	75.49	75.41	75.43
	S	65.87	65.26	65.42	65.33	65.41	65.02	64.75	64.42	64.06	64.09	64.96
9:1	**L**	83.32	83.38	83.41	83.44	83.46	83.47	83.5	83.51	83.53	83.54	83.45
	M	77.9	77.9	77.84	77.89	77.86	77.83	77.88	77.84	77.87	77.83	77.86
	S	70.36	70.05	69.93	69.79	70.14	69.84	70.13	70.16	70.23	70.16	70.08

(c) LR

% F-measure		DF Cutoff										
Train-Test Ratio	**category Size**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**overall**
1:9	**L**	81.38	81.38	81.36	81.36	81.35	81.35	81.35	81.34	81.33	81.33	81.35
	M	73.26	73.25	73.21	73.16	73.08	73.04	72.99	72.93	72.86	72.82	73.06
	S	49.41	49.49	49.49	49.41	49.31	49.2	49.04	48.77	48.5	48.21	49.08
1:1	**L**	82.62	82.6	82.6	82.6	82.6	82.6	82.6	82.6	82.59	82.59	82.6
	M	76.46	76.45	76.42	76.41	76.45	76.44	76.42	76.39	76.41	76.39	76.42
	S	64.33	64.44	64.49	64.46	64.49	64.64	64.29	64.29	64.21	63.9	64.35
9:1	**L**	83.71	83.7	83.7	83.7	83.71	83.7	83.69	83.7	83.7	83.7	83.7
	M	78.71	78.7	78.68	78.66	78.63	78.6	78.61	78.6	78.55	78.51	78.63
	S	69.06	69.16	69.27	69.37	69.28	69.31	69.31	69.27	69.27	69.32	69.26

Table B.9: Effect of Removing LFWs on Classification Performance for Different Category Sizes at DFC Levels 2 to 10

(a) MNB

Impact (% F-measure)		DF Cutoff									
Train-Test Ratio	category Size	2	3	4	5	6	7	8	9	10	overall
1:9	L	0.48	0.47	0.45	0.42	0.39	0.37	0.34	0.33	0.3	0.39
	M	5.99	7.9	8.89	9.5	9.89	10.13	10.35	10.48	10.58	9.3
	S	9.31	13.21	15.39	16.97	18.06	18.82	19.18	19.3	19.47	16.63
1:1	L	0.2	0.18	0.15	0.12	0.1	0.07	0.06	0.04	0.03	0.1
	M	4.31	5.34	5.59	5.7	5.74	5.79	5.77	5.82	5.77	5.54
	S	10.22	13.8	15.57	16.75	17.59	17.95	18.3	18.81	19.17	16.46
9:1	L	-0.06	-0.13	-0.18	-0.22	-0.24	-0.26	-0.29	-0.3	-0.31	-0.22
	M	3.81	4.26	4.45	4.47	4.5	4.52	4.48	4.47	4.44	4.38
	S	11.59	15.13	17.88	19.24	20.21	20.54	20.95	21.3	21.72	18.73

(b) SVM

Impact (% F-measure)		DF Cutoff									
Train-Test Ratio	category Size	2	3	4	5	6	7	8	9	10	overall
1:9	L	0.03	0.1	0.18	0.24	0.33	0.42	0.48	0.53	0.57	0.32
	M	-0.2	-0.36	-0.45	-0.59	-0.66	-0.71	-0.78	-0.86	-0.92	-0.61
	S	-0.57	-1.11	-1.36	-1.65	-2.03	-2.55	-2.99	-3.32	-3.86	-2.16
1:1	L	0.04	0.1	0.16	0.2	0.24	0.27	0.28	0.3	0.33	0.21
	M	-0.07	-0.1	-0.07	-0.04	-0.06	-0.03	-0.07	0.01	-0.08	-0.06
	S	-0.61	-0.44	-0.53	-0.46	-0.85	-1.11	-1.44	-1.8	-1.77	-1
9:1	L	0.06	0.09	0.12	0.15	0.15	0.18	0.19	0.21	0.22	0.15
	M	0	-0.06	-0.01	-0.04	-0.08	-0.02	-0.06	-0.04	-0.07	-0.04
	S	-0.31	-0.43	-0.57	-0.23	-0.53	-0.24	-0.21	-0.14	-0.2	-0.32

(c) LR

Impact (% F-measure)		DF Cutoff									
Train-Test Ratio	category Size	2	3	4	5	6	7	8	9	10	overall
1:9	L	0	-0.02	-0.02	-0.03	-0.03	-0.03	-0.04	-0.05	-0.05	-0.03
	M	0	-0.04	-0.09	-0.18	-0.22	-0.27	-0.33	-0.4	-0.43	-0.22
	S	0.08	0.08	-0.01	-0.11	-0.21	-0.38	-0.64	-0.91	-1.21	-0.37
1:1	L	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.03	-0.02
	M	0	-0.04	-0.05	-0.01	-0.02	-0.03	-0.06	-0.04	-0.06	-0.04
	S	0.1	0.16	0.12	0.15	0.31	-0.04	-0.05	-0.12	-0.43	0.02
9:1	L	-0.01	-0.01	0	0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
	M	-0.01	-0.03	-0.05	-0.08	-0.12	-0.1	-0.12	-0.16	-0.2	-0.1
	S	0.09	0.21	0.31	0.21	0.25	0.24	0.21	0.21	0.25	0.22

C. SUPPLEMENTARY TABLES FOR TYPE-M MAPPING AND TYPE-S GROUPING EXPERIMENTS

Table C.1: ANOVA Table (Impact): Coefficients of Classifiers as Indicator of Words' Predictive Categories for Type-M Mapping

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	55	0.246562	0.004483	10.65	<.0001
Error	754	0.317407	0.000421		
Corrected Total	809	0.563969			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Category (block)	29	0.108495	0.003741	8.89	<.0001
Classifier	2	0.134427	0.067213	159.67	<.0001
Coefficient	2	0.000181	9.05E-05	0.22	0.8066
Train-Test_Ratio	2	0.002216	0.001108	2.63	0.0726
Coefficient*Ratio	4	0.000389	9.72E-05	0.23	0.921
Classifier*Ratio	4	0.000357	8.92E-05	0.21	0.9319
Coefficient*Classifier	4	0.000271	6.78E-05	0.16	0.958
Coefficient*Classifier*Ratio	8	0.000227	2.84E-05	0.07	0.9998

Table C.2: ANOVA Table (Impact): Feature Selection Method for Selecting Top N Discriminatory Seedwords for Type-S Tagging

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	86	4.84583882	0.05634696	141.31	<.0001
Error	59313	23.65120506	0.00039875		
Corrected Total	59399	28.49704387			

Source	DF	Type III SS	Mean Square	F Value	Pr >F
Classifier	2	0.58603601	0.29301801	734.84	<.0001
Train-Test_Ratio	2	0.53726557	0.26863278	673.68	<.0001
Category (block)	29	3.13014905	0.10793617	270.68	<.0001
Feature_Selection	5	0.01293918	0.00258784	6.49	<.0001
TopN	4	0.00106292	0.00026573	0.67	0.6152
Feature_Selection*Classifier	10	0.01483461	0.00148346	3.72	<.0001
Classifier*Ratio	4	0.03742386	0.00935597	23.46	<.0001
Feature_Selection*Ratio	10	0.01328081	0.00132808	3.33	0.0002
Feature_Selection*Classifier*Ratio	20	0.01822002	0.000911	2.28	0.0009

Table C.3: Post Hoc Tests (Tukey and LSD) for Feature Selection Methods

Tukey Grouping		Impact Mean	N	Feature Selection	LSD Grouping	
	A	0.007479	9900	CHI		A
B	A	0.006684	9900	OR		B
B		0.006608	9900	LR	C	B
B		0.006095	9900	MNB	C	B
B		0.006035	9900	SVM		D
B		0.00591	9900	MI		D

D. SUPPLEMENTARY TABLES FOR FINAL EVALUATION

Table D.1: Overall Classification Performance: Grouping vs. Non-grouping

	Category Size	S			M			L			Regardless of Category Sizes			Overall
	Train-Test Ratio	1:9	1:1	9:1	1:9	1:1	9:1	1:9	1:1	9:1	1:9	1:1	9:1	
Non-Grouping StandardClassification	MNB	15.7%	27.9%	31.7%	56.5%	65.9%	68.1%	77.6%	79.3%	79.3%	42.9%	52.8%	55.5%	50.4%
	SVM	54.6%	68.8%	70.4%	72.3%	77.2%	77.9%	80.2%	83.0%	83.3%	66.3%	74.5%	75.5%	72.1%
	LR	49.4%	66.6%	69.1%	73.3%	78.0%	78.7%	81.4%	83.5%	83.7%	65.1%	74.2%	75.5%	71.6%
STEM	MNB	18.3%	29.7%	33.4%	58.8%	67.2%	68.9%	77.7%	79.0%	79.1%	42.9%	52.8%	55.5%	50.4%
	SVM	56.0%	69.3%	71.0%	72.4%	77.2%	77.9%	80.5%	83.0%	83.2%	66.3%	74.5%	75.5%	72.1%
	LR	51.3%	67.6%	70.6%	73.6%	78.0%	78.6%	81.5%	83.4%	83.6%	65.1%	74.2%	75.5%	71.6%
LEMMA	MNB	16.3%	28.0%	31.7%	57.3%	66.4%	68.3%	77.6%	79.2%	79.2%	43.6%	53.1%	55.6%	50.8%
	SVM	55.6%	69.3%	70.7%	72.5%	77.2%	78.0%	80.3%	83.0%	83.3%	66.8%	74.7%	75.7%	72.4%
	LR	50.1%	67.3%	70.7%	73.5%	78.1%	78.7%	81.4%	83.5%	83.7%	65.5%	74.5%	76.1%	72.0%
MS	MNB	20.3%	34.4%	37.7%	59.8%	68.4%	70.0%	77.5%	78.9%	79.0%	46.5%	56.6%	58.8%	54.0%
	SVM	56.2%	69.2%	70.4%	72.8%	77.3%	77.9%	80.1%	82.8%	83.1%	67.2%	74.7%	75.5%	72.5%
	LR	51.8%	67.3%	70.1%	73.9%	78.1%	78.7%	81.3%	83.3%	83.5%	66.3%	74.5%	75.9%	72.2%
MSS2	MNB	20.3%	34.1%	37.7%	62.4%	69.3%	70.2%	78.0%	79.1%	79.0%	48.0%	57.0%	58.8%	54.6%
	SVM	56.1%	69.7%	70.2%	73.1%	77.7%	78.0%	80.1%	83.0%	83.1%	67.4%	75.1%	75.5%	72.6%
	LR	51.9%	67.8%	70.1%	74.3%	78.5%	78.8%	81.4%	83.5%	83.5%	66.6%	74.9%	75.9%	72.5%
MSPA	MNB	21.2%	34.9%	38.7%	59.6%	68.3%	69.7%	77.5%	78.9%	78.9%	46.7%	56.8%	59.0%	54.1%
	SVM	58.6%	70.6%	71.6%	73.1%	77.3%	77.9%	80.1%	82.8%	83.2%	68.2%	75.2%	75.9%	73.1%
	LR	55.9%	70.1%	72.5%	74.5%	78.4%	78.9%	81.4%	83.4%	83.6%	68.2%	75.7%	76.9%	73.6%
MSS2PA	MNB	21.1%	34.6%	39.1%	61.9%	69.1%	70.0%	77.9%	79.0%	78.9%	48.0%	57.2%	59.3%	54.8%
	SVM	58.6%	71.3%	71.4%	73.3%	77.7%	78.0%	80.1%	83.1%	83.2%	68.3%	75.7%	75.9%	73.3%
	LR	56.0%	70.5%	72.6%	74.8%	78.8%	79.0%	81.4%	83.6%	83.6%	68.3%	76.1%	76.9%	73.8%
MS_NB	SVM	54.0%	64.6%	67.1%	70.8%	75.4%	76.8%	80.3%	82.4%	82.8%	65.3%	71.9%	73.6%	70.3%
	LR	56.5%	69.1%	70.6%	73.6%	77.6%	78.1%	81.4%	83.2%	83.4%	67.8%	74.9%	75.7%	72.8%
MSS2_NB	SVM	53.7%	65.1%	68.0%	70.5%	75.4%	77.0%	80.1%	82.6%	83.1%	65.0%	72.1%	74.1%	70.4%
	LR	56.5%	69.4%	70.5%	73.9%	78.0%	78.1%	81.3%	83.4%	83.4%	68.0%	75.2%	75.7%	73.0%
MSPA_NB	SVM	58.1%	67.3%	69.0%	71.6%	75.3%	77.0%	80.4%	82.5%	83.2%	67.2%	72.8%	74.5%	71.5%
	LR	60.6%	72.3%	72.5%	74.0%	77.9%	78.3%	81.5%	83.3%	83.5%	69.6%	76.2%	76.5%	74.1%
MSS2PA_NB	SVM	58.0%	67.3%	69.2%	71.2%	75.8%	77.0%	80.2%	82.8%	83.2%	67.0%	73.2%	74.5%	71.6%
	LR	60.7%	72.7%	72.7%	74.2%	78.2%	78.2%	81.4%	83.5%	83.5%	69.7%	76.6%	76.6%	74.3%

Table D.2: Category-wise Classification Performance for MNB at Train-Test Ratio of 1:9

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA
ANIMAL	81.0%	81.5%	81.2%	81.8%	81.8%	81.9%	81.9%
BICYCLE	68.3%	70.9%	69.6%	70.0%	80.7%	69.3%	79.6%
C289	42.6%	43.4%	43.0%	43.7%	44.0%	43.5%	43.8%
C289EYE	63.4%	64.9%	64.1%	65.8%	65.8%	64.8%	64.8%
C289FBEAR	39.6%	48.4%	42.5%	52.9%	52.7%	47.3%	47.1%
C289FBI	69.3%	72.2%	70.2%	72.9%	72.8%	72.6%	72.5%
C289FBNOSE	77.2%	81.5%	78.5%	84.8%	84.7%	82.0%	81.9%
CHOKING	33.6%	38.7%	34.0%	40.9%	40.6%	36.6%	36.5%
CUTTING	62.6%	63.0%	62.6%	63.1%	63.1%	62.8%	62.8%
DROWNING	21.1%	25.2%	21.8%	28.9%	28.9%	25.7%	25.7%
ELECTRICITY	36.7%	45.3%	39.0%	50.5%	50.5%	45.3%	45.3%
FALL	82.3%	82.4%	82.3%	82.3%	82.9%	82.3%	82.9%
FIREARM	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FIREFLAME	8.9%	9.5%	8.9%	11.4%	11.7%	10.5%	10.4%
HORSE	32.0%	39.4%	34.4%	39.1%	38.1%	38.1%	37.0%
HOTCOLDCOND	1.4%	1.3%	1.1%	2.6%	2.6%	2.2%	2.1%
HOTOBJ	81.5%	81.8%	82.0%	82.0%	81.9%	81.4%	81.4%
MACHINERY	56.0%	56.4%	56.2%	56.8%	56.8%	56.1%	56.1%
MOTORCYCLE	78.4%	79.6%	79.0%	80.6%	83.6%	79.6%	82.7%
MOTORVEHICLE	87.7%	87.8%	87.8%	87.9%	88.8%	87.9%	88.7%
OTHERTRANSPORT	4.8%	6.5%	5.4%	7.0%	6.9%	6.9%	6.7%
PA_ALCOHOL	1.0%	1.3%	1.0%	1.6%	1.7%	5.3%	5.2%
PA_CHEMICAL	37.3%	41.6%	38.4%	47.1%	46.9%	54.0%	53.8%
PA_DRUG	74.0%	74.4%	74.4%	75.9%	76.0%	78.6%	78.6%
PA_DRUGALCO	24.4%	24.0%	24.3%	25.6%	25.6%	35.5%	35.4%
PA_FOOD	25.9%	33.2%	29.3%	36.9%	36.8%	47.0%	46.7%
PA_OTHERS	20.0%	22.9%	19.7%	24.8%	24.8%	25.0%	24.9%
PA_PLANT	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PEDESTRIAN	3.9%	5.7%	4.1%	5.9%	35.9%	5.9%	32.5%
STRUCKCOLLISION	72.9%	72.9%	72.9%	72.8%	73.0%	72.6%	72.8%

Table D.3: Category-wise Classification Performance for SVM at Train-Test Ratio of 1:9

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA	MS_NB	MSS2_NB	MSPA_NB	MSS2PA_NB
ANIMAL	88.4%	88.4%	88.6%	88.7%	88.7%	88.5%	88.6%	88.2%	88.1%	88.2%	88.0%
BICYCLE	87.2%	87.5%	87.2%	87.3%	87.9%	87.1%	87.6%	86.2%	84.7%	86.1%	84.3%
C289	47.0%	47.3%	47.2%	46.8%	46.8%	46.8%	46.8%	46.7%	46.2%	46.7%	46.3%
C289EYE	69.2%	69.5%	69.5%	69.3%	69.4%	69.1%	69.1%	66.0%	65.8%	65.8%	65.6%
C289FBEAR	71.8%	72.1%	72.3%	72.5%	72.5%	72.0%	71.9%	68.7%	68.6%	68.1%	68.1%
C289FBI	75.3%	75.6%	75.5%	76.5%	76.5%	76.3%	76.3%	73.3%	73.2%	74.0%	73.9%
C289FBNOSE	92.1%	92.3%	92.2%	92.5%	92.5%	92.4%	92.4%	90.4%	90.5%	90.5%	90.4%
CHOKING	83.2%	85.8%	83.2%	84.0%	84.0%	83.5%	83.4%	81.9%	81.4%	81.5%	81.0%
CUTTING	62.4%	62.9%	62.5%	62.7%	62.7%	62.4%	62.4%	62.6%	62.6%	62.4%	62.2%
DROWNING	72.4%	72.9%	72.9%	74.3%	74.4%	73.4%	73.6%	70.8%	70.2%	72.7%	72.5%
ELECTRICITY	85.9%	87.0%	86.1%	86.6%	86.7%	86.3%	86.3%	84.6%	84.0%	84.5%	84.3%
FALL	85.3%	85.6%	85.4%	85.3%	85.3%	85.3%	85.3%	85.5%	85.2%	85.7%	85.3%
FIREARM	20.0%	20.4%	20.8%	20.1%	19.1%	19.4%	18.7%	19.2%	17.8%	21.0%	20.9%
FIREFLAME	57.6%	57.3%	58.0%	58.2%	58.1%	57.4%	57.5%	56.6%	56.7%	56.2%	56.2%
HORSE	92.3%	92.0%	92.4%	92.4%	92.4%	92.2%	92.2%	82.7%	84.8%	90.0%	85.7%
HOTCOLDCOND	60.8%	60.7%	60.9%	64.6%	64.5%	64.1%	64.0%	61.5%	60.7%	61.4%	61.4%
HOTOBJ	85.2%	84.9%	85.3%	85.3%	85.4%	85.1%	85.2%	84.2%	84.1%	84.2%	84.2%
MACHINERY	66.3%	66.7%	66.4%	66.4%	66.4%	66.2%	66.2%	65.7%	65.6%	65.6%	65.4%
MOTORCYCLE	87.1%	87.2%	87.2%	87.5%	88.0%	87.3%	87.7%	86.6%	80.6%	86.5%	85.3%
MOTORVEHICLE	89.6%	89.5%	89.7%	90.2%	90.4%	90.1%	90.3%	89.5%	88.7%	89.4%	89.0%
OTHERTRANSPORT	26.1%	26.3%	26.4%	26.3%	26.2%	26.2%	25.8%	26.4%	25.9%	26.4%	26.0%
PA_ALCOHOL	36.9%	38.1%	37.4%	40.0%	40.1%	47.9%	48.0%	36.8%	37.6%	46.1%	46.1%
PA_CHEMICAL	52.6%	52.8%	52.9%	54.7%	54.6%	59.1%	59.1%	52.3%	52.1%	55.9%	55.8%
PA_DRUG	78.4%	79.0%	78.9%	80.2%	80.2%	83.0%	83.0%	78.2%	78.2%	82.5%	82.5%
PA_DRUGALCO	60.5%	61.6%	61.3%	61.9%	62.1%	70.5%	70.5%	55.3%	55.2%	71.7%	71.5%
PA_FOOD	58.1%	61.0%	60.7%	60.4%	60.4%	67.7%	67.7%	58.8%	58.6%	64.4%	64.5%
PA_OTHERS	47.1%	47.8%	48.0%	48.6%	48.5%	50.7%	50.6%	46.5%	46.4%	49.0%	49.0%
PA_PLANT	18.3%	22.9%	21.9%	19.8%	19.2%	24.0%	24.1%	22.5%	22.6%	30.5%	30.8%
PEDESTRIAN	58.2%	57.2%	58.1%	58.9%	62.6%	58.5%	60.9%	55.3%	58.4%	54.7%	57.5%
STRUCKCOLLISION	75.0%	75.3%	75.1%	74.8%	74.9%	74.8%	74.9%	75.1%	75.0%	75.2%	75.1%

Table D.4: Category-wise Classification Performance for LR at Train-Test Ratio of 1:9

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA	MS_NB	MSS2_NB	MSPA_NB	MSS2PA_NB
ANIMAL	88.9%	88.9%	89.1%	89.3%	89.2%	89.2%	89.2%	89.8%	89.7%	89.8%	89.7%
BICYCLE	88.0%	88.2%	88.0%	88.0%	88.6%	87.9%	88.4%	87.8%	86.4%	87.7%	85.9%
C289	49.8%	49.9%	49.9%	49.7%	49.8%	49.9%	49.9%	49.2%	48.9%	49.4%	49.0%
C289EYE	73.2%	73.7%	73.7%	73.4%	73.4%	73.2%	73.2%	70.4%	70.2%	70.2%	70.1%
C289FBEAR	74.6%	75.4%	75.6%	74.9%	75.0%	74.6%	74.7%	71.6%	71.7%	71.0%	71.1%
C289FBI	76.2%	77.4%	76.6%	77.6%	77.6%	77.9%	77.8%	77.4%	77.2%	77.6%	77.6%
C289FBNOSE	93.0%	93.2%	93.1%	93.6%	93.6%	93.5%	93.5%	92.8%	92.8%	92.7%	92.7%
CHOKING	80.7%	87.4%	80.7%	82.3%	82.3%	81.5%	81.5%	86.1%	85.9%	85.7%	85.6%
CUTTING	65.2%	65.5%	65.2%	65.5%	65.5%	65.3%	65.3%	65.2%	65.2%	65.0%	64.9%
DROWNING	65.6%	66.7%	65.8%	65.8%	65.8%	65.5%	65.5%	75.8%	75.9%	75.6%	75.5%
ELECTRICITY	82.8%	86.2%	83.1%	86.3%	86.2%	86.1%	86.2%	87.7%	87.6%	87.6%	87.5%
FALL	86.4%	86.5%	86.4%	86.3%	86.4%	86.4%	86.5%	86.5%	86.3%	86.6%	86.4%
FIREARM	1.2%	1.5%	1.2%	1.5%	1.5%	1.8%	1.8%	15.7%	15.2%	13.7%	14.0%
FIREFLAME	57.8%	58.7%	58.8%	58.9%	58.6%	58.5%	58.5%	61.7%	61.5%	61.4%	61.4%
HORSE	92.4%	92.4%	92.6%	92.6%	92.5%	92.5%	92.5%	92.6%	92.6%	92.5%	92.6%
HOTCOLDCOND	49.0%	49.2%	48.6%	58.3%	58.3%	56.5%	56.6%	67.5%	67.7%	67.1%	67.0%
HOTOBJ	86.4%	86.5%	86.6%	86.6%	86.6%	86.5%	86.5%	86.8%	86.8%	87.0%	87.0%
MACHINERY	67.9%	68.2%	68.0%	68.2%	68.2%	68.1%	68.1%	68.2%	68.1%	68.2%	68.1%
MOTORCYCLE	88.0%	88.1%	88.1%	88.6%	89.0%	88.5%	88.8%	88.8%	88.6%	88.8%	88.3%
MOTORVEHICLE	90.9%	90.8%	90.9%	91.7%	91.8%	91.6%	91.7%	91.6%	91.6%	91.5%	91.5%
OTHERTRANSPORT	22.5%	23.0%	22.7%	22.9%	22.9%	22.6%	22.7%	26.6%	26.4%	26.6%	26.2%
PA_ALCOHOL	32.8%	34.5%	33.3%	35.7%	36.4%	48.4%	48.5%	34.3%	34.6%	49.5%	49.4%
PA_CHEMICAL	52.4%	53.5%	52.8%	55.2%	55.2%	63.0%	63.0%	56.9%	56.8%	61.0%	61.0%
PA_DRUG	78.7%	79.3%	79.3%	81.1%	81.1%	84.9%	84.9%	81.1%	81.1%	85.4%	85.4%
PA_DRUGALCO	63.3%	63.6%	64.3%	64.6%	65.0%	74.3%	74.3%	61.0%	60.8%	76.1%	76.1%
PA_FOOD	57.7%	61.0%	60.8%	60.8%	61.0%	72.5%	72.6%	63.8%	64.0%	70.7%	70.7%
PA_OTHERS	47.2%	48.0%	47.8%	48.7%	48.8%	52.0%	52.1%	50.0%	50.0%	53.0%	52.8%
PA_PLANT	5.5%	7.5%	7.0%	6.9%	6.9%	17.9%	17.9%	17.5%	17.7%	26.5%	27.2%
PEDESTRIAN	57.3%	56.7%	57.0%	57.9%	63.3%	57.6%	61.3%	54.6%	61.7%	54.3%	60.4%
STRUCKCOLLISION	76.4%	76.4%	76.4%	76.3%	76.4%	76.4%	76.4%	76.4%	76.4%	76.4%	76.4%

Table D.5: Category-wise Classification Performance for MNB at Train-Test Ratio of 1:1

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA
ANIMAL	84.1%	84.2%	84.1%	84.8%	84.8%	84.7%	84.8%
BICYCLE	79.1%	80.3%	79.8%	80.0%	80.6%	79.7%	80.3%
C289	48.4%	48.1%	48.4%	48.4%	49.1%	48.3%	48.9%
C289EYE	68.8%	69.0%	69.0%	69.4%	69.5%	68.8%	68.9%
C289FBEAR	63.8%	68.7%	66.0%	71.5%	71.4%	70.3%	70.3%
C289FBI	78.9%	79.7%	79.1%	80.5%	80.6%	80.3%	80.4%
C289FBNOSE	90.0%	91.3%	90.3%	92.9%	93.0%	92.6%	92.7%
CHOKING	54.2%	58.5%	54.3%	61.2%	61.0%	58.7%	58.4%
CUTTING	65.5%	65.4%	65.5%	65.4%	65.6%	65.2%	65.4%
DROWNING	48.4%	52.3%	49.9%	61.0%	60.5%	58.5%	58.0%
ELECTRICITY	63.8%	69.3%	65.5%	77.3%	77.3%	75.6%	75.5%
FALL	83.9%	83.8%	83.9%	83.7%	83.9%	83.7%	83.8%
FIREARM	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FIREFLAME	21.2%	22.3%	20.3%	27.9%	27.3%	26.1%	25.9%
HORSE	54.9%	61.3%	56.7%	63.1%	63.1%	62.2%	62.2%
HOTCOLDCOND	5.1%	3.1%	4.6%	8.9%	8.5%	6.8%	6.8%
HOTOBJ	83.7%	83.6%	83.8%	83.9%	84.0%	83.6%	83.7%
MACHINERY	59.1%	58.9%	59.1%	59.5%	59.8%	58.9%	59.2%
MOTORCYCLE	84.2%	84.7%	84.6%	84.7%	85.3%	84.5%	85.1%
MOTORVEHICLE	88.6%	88.7%	88.7%	88.9%	89.3%	88.8%	89.2%
OTHERTRANSPORT	16.9%	19.5%	17.8%	21.9%	22.4%	22.4%	23.0%
PA_ALCOHOL	1.3%	1.7%	1.1%	3.1%	3.1%	11.6%	10.7%
PA_CHEMICAL	59.8%	61.4%	60.2%	64.0%	64.3%	67.0%	67.4%
PA_DRUG	77.8%	77.9%	78.0%	79.1%	79.1%	80.5%	80.5%
PA_DRUGALCO	25.2%	25.9%	24.6%	27.2%	26.0%	36.2%	35.2%
PA_FOOD	51.9%	56.3%	53.0%	63.6%	63.6%	68.2%	68.3%
PA_OTHERS	36.0%	37.1%	34.6%	47.8%	47.7%	41.8%	42.1%
PA_PLANT	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PEDESTRIAN	16.2%	19.5%	16.9%	25.0%	36.0%	24.1%	33.7%
STRUCKCOLLISION	74.6%	74.3%	74.5%	74.2%	74.3%	74.1%	74.2%

Table D.6: Category-wise Classification Performance for SVM at Train-Test Ratio of 1:1

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA	MS_NB	MSS2_NB	MSPA_NB	MSS2PA_NB
ANIMAL	91.0%	91.0%	91.1%	91.0%	91.1%	91.0%	91.1%	90.5%	90.6%	90.3%	90.4%
BICYCLE	90.0%	90.2%	90.1%	89.7%	89.9%	89.7%	89.9%	88.9%	89.1%	89.2%	89.1%
C289	53.8%	53.6%	53.9%	53.3%	54.4%	53.5%	54.5%	52.8%	53.9%	53.0%	54.1%
C289EYE	74.0%	74.5%	74.4%	73.9%	74.4%	73.8%	74.3%	71.6%	72.0%	71.5%	72.1%
C289FBEAR	76.3%	76.4%	76.5%	76.8%	77.1%	76.4%	76.7%	70.7%	71.0%	73.0%	73.4%
C289FBI	81.5%	81.7%	81.7%	82.0%	82.4%	81.5%	82.1%	79.6%	79.5%	79.7%	79.8%
C289FBNOSE	93.9%	93.9%	93.9%	94.1%	94.3%	93.9%	94.1%	87.0%	82.3%	80.6%	84.7%
CHOKING	89.9%	90.4%	89.6%	90.1%	90.6%	90.0%	90.4%	88.2%	87.9%	87.8%	87.9%
CUTTING	68.1%	68.5%	68.2%	68.0%	68.5%	68.1%	68.6%	67.3%	67.9%	67.4%	67.8%
DROWNING	81.8%	81.2%	82.0%	82.3%	82.4%	81.7%	81.6%	80.7%	80.4%	80.8%	80.8%
ELECTRICITY	91.6%	91.1%	91.5%	90.7%	90.9%	91.0%	91.2%	89.5%	89.5%	89.7%	89.7%
FALL	87.5%	87.5%	87.5%	87.3%	87.5%	87.4%	87.5%	86.9%	87.0%	87.1%	87.4%
FIREARM	55.5%	53.3%	55.2%	57.9%	55.9%	54.4%	53.9%	55.3%	54.5%	55.2%	54.8%
FIREFLAME	65.7%	65.8%	65.5%	65.0%	65.3%	64.7%	65.1%	63.6%	59.0%	63.0%	63.6%
HORSE	93.5%	92.9%	93.6%	93.6%	93.7%	93.5%	93.6%	92.5%	92.7%	92.4%	92.6%
HOTCOLDCOND	73.9%	74.2%	74.1%	74.8%	76.6%	74.4%	75.9%	70.8%	71.1%	70.4%	71.5%
HOTOBJ	87.8%	87.8%	88.0%	87.7%	88.0%	87.7%	87.9%	87.1%	87.0%	86.9%	86.9%
MACHINERY	71.6%	71.8%	71.6%	71.4%	71.9%	71.5%	72.0%	70.6%	71.0%	70.5%	71.1%
MOTORCYCLE	90.2%	90.2%	90.2%	90.3%	90.5%	90.2%	90.4%	89.7%	89.8%	89.7%	89.7%
MOTORVEHICLE	92.2%	92.3%	92.2%	92.2%	92.4%	92.1%	92.3%	92.0%	92.0%	91.9%	91.9%
OTHERTRANSPORT	32.8%	32.9%	33.2%	32.8%	33.2%	32.9%	33.6%	31.5%	32.0%	31.6%	32.0%
PA_ALCOHOL	57.6%	59.3%	58.1%	60.1%	60.6%	65.3%	66.1%	52.2%	55.4%	61.4%	62.7%
PA_CHEMICAL	64.8%	64.6%	64.2%	65.2%	66.2%	66.7%	67.6%	62.3%	62.8%	63.7%	64.7%
PA_DRUG	85.0%	84.9%	85.0%	85.6%	86.0%	86.7%	87.1%	83.4%	83.5%	85.3%	85.6%
PA_DRUGALCO	71.8%	72.0%	71.8%	71.6%	72.3%	76.7%	77.4%	54.2%	61.1%	60.1%	54.2%
PA_FOOD	71.0%	72.6%	72.4%	69.4%	70.3%	73.6%	74.3%	66.9%	67.5%	70.8%	71.5%
PA_OTHERS	55.8%	56.4%	56.2%	56.6%	57.4%	57.7%	58.5%	49.4%	50.0%	55.3%	56.1%
PA_PLANT	42.4%	45.5%	45.7%	42.2%	44.2%	47.5%	49.9%	39.2%	40.0%	45.4%	47.1%
PEDESTRIAN	65.2%	64.4%	65.0%	65.6%	66.5%	65.3%	65.9%	63.9%	63.8%	63.6%	63.3%
STRUCKCOLLISION	78.5%	78.5%	78.5%	78.2%	78.5%	78.3%	78.6%	77.9%	78.2%	77.9%	78.2%

Table D.7: Category-wise Classification Performance for LR at Train-Test Ratio of 1:1

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA	MS_NB	MSS2_NB	MSPA_NB	MSS2PA_NB
ANIMAL	91.2%	91.2%	91.3%	91.2%	91.3%	91.3%	91.3%	91.4%	91.5%	91.4%	91.5%
BICYCLE	90.3%	90.4%	90.3%	90.1%	90.3%	90.1%	90.3%	89.9%	89.9%	89.9%	89.9%
C289	55.4%	55.0%	55.4%	55.0%	56.0%	55.2%	56.3%	54.6%	55.7%	54.8%	55.9%
C289EYE	76.0%	76.2%	76.3%	76.0%	76.4%	76.0%	76.3%	74.1%	74.5%	74.1%	74.5%
C289FBEAR	78.1%	78.6%	78.5%	78.2%	78.5%	78.1%	78.3%	76.0%	76.5%	75.7%	76.2%
C289FBI	83.1%	83.5%	83.4%	83.3%	83.8%	83.2%	83.8%	83.1%	83.5%	83.0%	83.4%
C289FBNOSE	94.8%	94.8%	94.8%	95.1%	95.2%	95.0%	95.1%	94.6%	94.8%	94.6%	94.7%
CHOKING	90.4%	91.6%	90.6%	91.3%	91.5%	91.1%	91.4%	90.7%	90.8%	90.3%	90.6%
CUTTING	69.3%	69.3%	69.3%	69.2%	69.6%	69.3%	69.8%	68.8%	69.2%	68.9%	69.4%
DROWNING	82.2%	82.1%	82.2%	82.4%	82.7%	82.6%	82.7%	83.4%	83.6%	83.7%	83.8%
ELECTRICITY	90.6%	91.0%	90.7%	91.0%	91.1%	91.1%	91.2%	91.5%	91.7%	91.7%	91.8%
FALL	87.9%	87.8%	87.9%	87.7%	87.9%	87.8%	88.0%	87.7%	87.9%	87.9%	88.0%
FIREARM	34.4%	36.0%	35.7%	35.1%	35.6%	34.4%	33.2%	53.0%	50.9%	54.5%	53.7%
FIREFLAME	67.3%	67.0%	67.1%	67.2%	67.3%	66.5%	66.8%	67.4%	67.7%	67.5%	67.9%
HORSE	93.4%	93.0%	93.6%	93.6%	93.7%	93.6%	93.7%	93.4%	93.5%	93.4%	93.5%
HOTCOLDCOND	74.6%	74.2%	74.3%	76.4%	77.5%	76.3%	77.4%	76.0%	76.6%	75.8%	76.4%
HOTOBJ	88.7%	88.6%	88.8%	88.6%	88.8%	88.5%	88.7%	88.7%	88.9%	88.8%	89.0%
MACHINERY	72.3%	72.4%	72.3%	72.3%	72.7%	72.3%	72.7%	71.9%	72.4%	72.0%	72.5%
MOTORCYCLE	90.7%	90.7%	90.7%	90.9%	91.2%	91.0%	91.2%	91.0%	91.0%	91.0%	91.1%
MOTORVEHICLE	92.9%	92.8%	92.8%	92.9%	93.0%	92.9%	93.0%	92.9%	92.9%	92.9%	92.9%
OTHERTRANSPORT	32.6%	32.5%	32.6%	32.3%	33.2%	32.7%	33.5%	32.2%	32.8%	32.3%	32.8%
PA_ALCOHOL	56.1%	56.1%	55.8%	56.7%	58.0%	65.7%	66.2%	56.4%	57.1%	65.7%	66.5%
PA_CHEMICAL	66.4%	66.3%	66.4%	67.1%	68.1%	69.9%	70.8%	66.6%	67.7%	68.6%	69.7%
PA_DRUG	85.7%	85.7%	85.8%	86.6%	86.9%	88.2%	88.6%	86.2%	86.5%	88.3%	88.7%
PA_DRUGALCO	72.8%	73.4%	73.1%	74.0%	74.4%	79.9%	80.9%	69.2%	69.5%	80.2%	80.7%
PA_FOOD	72.6%	74.6%	74.4%	72.9%	73.7%	77.6%	78.4%	72.6%	73.2%	76.0%	76.6%
PA_OTHERS	56.5%	57.0%	56.7%	57.4%	57.9%	59.1%	59.5%	56.6%	57.3%	58.6%	59.2%
PA_PLANT	34.9%	40.0%	39.8%	35.3%	36.2%	46.1%	48.1%	43.4%	45.5%	51.0%	52.4%
PEDESTRIAN	65.8%	64.9%	65.8%	65.9%	66.5%	65.9%	66.5%	64.3%	64.2%	64.2%	64.2%
STRUCKCOLLISION	79.0%	78.9%	79.0%	78.8%	79.1%	78.9%	79.2%	78.7%	78.9%	78.8%	79.0%

Table D.8: Category-wise Classification Performance for MNB at Train-Test Ratio of 9:1

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA
ANIMAL	84.8%	84.8%	84.7%	85.1%	85.2%	85.1%	85.1%
BICYCLE	80.8%	81.5%	81.2%	81.3%	81.4%	80.7%	80.7%
C289	48.8%	48.3%	48.7%	48.9%	48.8%	48.6%	48.5%
C289EYE	69.2%	69.2%	69.3%	69.4%	69.4%	69.1%	69.2%
C289FBEAR	70.0%	71.6%	70.4%	73.1%	73.1%	72.2%	72.5%
C289FBI	80.7%	81.1%	80.7%	81.0%	81.0%	80.6%	80.7%
C289FBNOSE	91.8%	92.6%	91.9%	93.3%	93.3%	93.0%	93.1%
CHOKING	60.3%	64.7%	60.9%	68.5%	68.4%	66.8%	66.9%
CUTTING	65.6%	65.6%	65.6%	65.5%	65.5%	65.3%	65.3%
DROWNING	56.4%	59.6%	57.4%	68.5%	68.5%	65.5%	66.5%
ELECTRICITY	72.1%	75.6%	72.5%	79.3%	79.2%	76.8%	76.9%
FALL	84.0%	83.9%	84.0%	83.9%	83.9%	83.7%	83.7%
FIREARM	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FIREFLAME	26.7%	28.8%	25.5%	36.7%	36.4%	34.6%	35.3%
HORSE	62.3%	67.8%	63.9%	70.2%	70.1%	68.2%	68.8%
HOTCOLDCOND	6.5%	4.6%	5.8%	12.2%	11.9%	9.6%	9.3%
HOTOBJ	84.2%	84.0%	84.2%	84.4%	84.4%	84.1%	84.1%
MACHINERY	59.1%	58.7%	59.0%	58.9%	59.0%	58.9%	58.9%
MOTORCYCLE	85.2%	85.3%	85.4%	85.5%	85.5%	85.3%	85.3%
MOTORVEHICLE	88.8%	89.0%	88.9%	89.1%	89.2%	89.0%	89.2%
OTHERTRANSPORT	21.8%	23.1%	22.1%	24.7%	24.7%	24.7%	25.0%
PA_ALCOHOL	1.4%	3.2%	2.2%	4.9%	5.2%	15.7%	17.3%
PA_CHEMICAL	62.7%	63.4%	62.4%	65.3%	65.5%	67.8%	68.0%
PA_DRUG	78.3%	78.1%	78.2%	79.5%	79.6%	80.5%	80.6%
PA_DRUGALCO	25.4%	25.7%	25.1%	28.2%	27.9%	38.3%	38.2%
PA_FOOD	58.0%	63.2%	59.2%	69.0%	69.1%	71.8%	72.7%
PA_OTHERS	41.4%	42.5%	39.6%	47.8%	47.9%	46.6%	47.2%
PA_PLANT	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
PEDESTRIAN	23.8%	27.4%	24.2%	34.7%	37.2%	32.2%	35.2%
STRUCKCOLLISION	74.5%	74.2%	74.5%	74.2%	74.2%	74.0%	74.0%

Table D.9: Category-wise Classification Performance for SVM at Train-Test Ratio of 9:1

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA	MS_NB	MSS2_NB	MSPA_NB	MSS2PA_NB
ANIMAL	91.4%	91.4%	91.4%	91.5%	91.5%	91.4%	91.4%	90.8%	91.1%	91.1%	91.1%
BICYCLE	90.4%	90.6%	90.5%	90.4%	90.4%	90.3%	90.3%	89.9%	90.2%	90.1%	90.1%
C289	54.3%	54.1%	54.4%	53.9%	53.9%	54.0%	54.0%	53.4%	53.6%	53.8%	53.7%
C289EYE	75.0%	75.4%	75.3%	75.2%	75.2%	75.1%	74.9%	73.0%	73.2%	73.1%	73.0%
C289FBEAR	77.1%	77.8%	77.5%	77.5%	77.7%	77.5%	77.5%	75.7%	75.7%	75.7%	75.4%
C289FBI	83.1%	83.1%	83.3%	82.8%	82.9%	82.6%	82.7%	81.4%	81.5%	81.1%	81.3%
C289FBNOSE	94.6%	94.5%	94.5%	94.7%	94.6%	94.6%	94.6%	93.4%	93.4%	93.4%	93.4%
CHOKING	90.6%	90.4%	90.4%	90.4%	90.3%	90.2%	90.2%	89.3%	89.4%	88.9%	88.9%
CUTTING	68.4%	68.9%	68.6%	68.4%	68.4%	68.4%	68.4%	68.0%	68.1%	68.1%	68.2%
DROWNING	83.8%	83.9%	83.8%	84.0%	83.8%	83.8%	83.5%	82.2%	82.7%	82.5%	82.4%
ELECTRICITY	92.2%	92.6%	92.1%	90.1%	90.2%	90.4%	90.3%	89.1%	89.6%	89.5%	89.5%
FALL	87.7%	87.7%	87.8%	87.5%	87.5%	87.6%	87.6%	87.1%	87.6%	87.7%	87.7%
FIREARM	56.5%	57.0%	58.7%	56.9%	56.9%	54.1%	54.1%	57.6%	57.4%	49.7%	52.2%
FIREFLAME	65.9%	66.8%	66.5%	66.2%	65.6%	65.1%	64.6%	63.7%	63.5%	63.4%	63.1%
HORSE	93.7%	93.3%	93.8%	93.9%	93.8%	93.9%	93.9%	93.6%	93.9%	93.7%	93.8%
HOTCOLDCOND	76.0%	76.6%	75.8%	77.2%	77.3%	76.9%	76.9%	73.6%	73.4%	72.5%	73.2%
HOTOBJ	88.3%	88.6%	88.5%	88.3%	88.3%	88.1%	88.2%	87.7%	87.9%	87.8%	87.9%
MACHINERY	71.9%	72.2%	72.0%	71.7%	71.7%	71.8%	71.7%	71.1%	71.2%	71.3%	71.3%
MOTORCYCLE	90.7%	90.6%	90.6%	90.6%	90.6%	90.6%	90.6%	90.2%	90.5%	90.3%	90.4%
MOTORVEHICLE	92.5%	92.7%	92.6%	92.6%	92.7%	92.6%	92.6%	92.3%	92.4%	92.4%	92.4%
OTHERTRANSPORT	32.8%	33.2%	33.4%	33.0%	32.9%	32.7%	32.8%	31.4%	31.5%	31.7%	31.5%
PA_ALCOHOL	62.2%	62.5%	60.7%	63.4%	62.5%	69.2%	68.8%	56.7%	58.9%	64.1%	63.5%
PA_CHEMICAL	66.9%	66.1%	66.3%	67.3%	66.9%	67.2%	67.6%	64.2%	64.4%	65.0%	65.1%
PA_DRUG	86.1%	85.9%	86.1%	86.3%	86.4%	87.0%	87.2%	83.9%	84.4%	86.1%	86.1%
PA_DRUGALCO	72.9%	72.2%	72.0%	72.4%	72.0%	75.9%	76.4%	59.6%	65.6%	77.2%	76.6%
PA_FOOD	73.5%	74.0%	74.6%	73.9%	73.7%	75.1%	74.3%	70.8%	71.3%	70.7%	70.6%
PA_OTHERS	58.7%	58.7%	59.0%	58.8%	58.5%	58.6%	58.4%	56.2%	56.1%	56.9%	56.8%
PA_PLANT	41.8%	45.8%	43.8%	41.3%	41.8%	48.1%	47.7%	39.7%	40.1%	43.8%	43.9%
PEDESTRIAN	66.9%	66.0%	67.1%	67.0%	67.5%	66.9%	67.3%	65.2%	65.7%	64.6%	64.8%
STRUCKCOLLISION	78.9%	78.8%	78.9%	78.7%	78.6%	78.7%	78.7%	78.5%	78.6%	78.7%	78.7%

Table D.10: Category-wise Classification Performance for LR at Train-Test Ratio of 9:1

Category	control	STEM	LEMMA	MS	MSS2	MSPA	MSS2PA	MS_NB	MSS2_NB	MSPA_NB	MSS2PA_NB
ANIMAL	91.7%	91.6%	91.7%	91.6%	91.6%	91.7%	91.6%	91.6%	91.6%	91.7%	91.7%
BICYCLE	90.7%	90.7%	90.7%	90.6%	90.6%	90.6%	90.6%	90.5%	90.5%	90.5%	90.4%
C289	55.7%	55.3%	55.6%	55.3%	55.4%	55.6%	55.6%	54.9%	55.0%	55.0%	55.0%
C289EYE	76.5%	76.4%	76.6%	76.4%	76.3%	76.6%	76.4%	74.9%	74.9%	74.8%	74.7%
C289FBEAR	78.8%	79.1%	79.0%	78.9%	78.9%	78.9%	78.7%	77.1%	77.1%	76.8%	76.8%
C289FBI	84.4%	84.4%	84.5%	84.3%	84.3%	84.1%	84.1%	83.8%	83.7%	83.5%	83.6%
C289FBNOSE	95.2%	95.2%	95.0%	95.4%	95.4%	95.4%	95.4%	94.9%	95.0%	94.8%	94.9%
CHOKING	91.5%	91.9%	91.7%	91.8%	91.9%	91.8%	91.8%	91.1%	91.1%	90.9%	90.8%
CUTTING	69.5%	69.4%	69.4%	69.2%	69.2%	69.3%	69.4%	68.9%	68.8%	69.0%	69.0%
DROWNING	84.7%	84.3%	85.1%	85.2%	85.2%	85.2%	85.0%	84.9%	84.8%	84.8%	84.9%
ELECTRICITY	91.4%	92.0%	91.6%	90.8%	90.8%	91.3%	91.0%	91.5%	91.5%	91.3%	91.3%
FALL	88.1%	88.0%	88.1%	87.8%	87.8%	88.0%	87.9%	87.8%	87.8%	87.9%	87.9%
FIREARM	40.7%	48.6%	49.0%	47.4%	47.4%	43.7%	45.4%	57.9%	57.9%	53.3%	56.4%
FIREFLAME	68.6%	68.7%	68.7%	68.2%	68.1%	67.9%	67.7%	67.6%	67.2%	66.8%	66.3%
HORSE	93.7%	93.2%	93.9%	93.9%	93.9%	93.9%	93.9%	93.9%	93.9%	94.0%	93.9%
HOTCOLDCOND	77.0%	77.0%	77.4%	77.8%	77.8%	77.5%	78.2%	77.0%	77.0%	76.4%	76.8%
HOTOBJ	89.1%	88.9%	89.1%	89.1%	89.3%	89.0%	89.1%	88.9%	89.0%	88.8%	89.0%
MACHINERY	72.6%	72.6%	72.5%	72.4%	72.4%	72.6%	72.5%	72.2%	72.2%	72.2%	72.2%
MOTORCYCLE	91.0%	91.0%	91.0%	91.2%	91.3%	91.1%	91.1%	91.4%	91.4%	91.3%	91.2%
MOTORVEHICLE	93.1%	92.9%	93.1%	93.1%	93.1%	93.0%	93.0%	93.1%	93.1%	93.1%	93.1%
OTHERTRANSPORT	33.7%	33.3%	33.7%	33.4%	33.3%	33.4%	33.3%	32.0%	31.9%	32.0%	31.9%
PA_ALCOHOL	60.1%	61.9%	60.8%	61.4%	61.4%	70.0%	69.9%	60.6%	60.9%	67.3%	67.0%
PA_CHEMICAL	68.2%	67.8%	67.9%	68.9%	68.6%	70.7%	71.1%	67.4%	67.3%	68.7%	68.7%
PA_DRUG	86.9%	86.9%	87.0%	87.3%	87.4%	88.7%	88.9%	86.3%	86.5%	88.7%	88.8%
PA_DRUGALCO	73.5%	74.8%	74.5%	74.8%	74.8%	81.2%	81.5%	67.2%	67.0%	80.4%	80.3%
PA_FOOD	75.5%	76.0%	76.7%	75.7%	75.6%	78.6%	78.3%	74.8%	74.7%	76.3%	76.3%
PA_OTHERS	58.7%	59.0%	58.7%	59.3%	59.4%	60.3%	60.2%	58.2%	58.1%	59.3%	59.2%
PA_PLANT	38.0%	42.6%	43.3%	38.8%	38.8%	50.2%	49.8%	45.3%	45.6%	50.3%	50.9%
PEDESTRIAN	67.7%	66.8%	67.6%	67.7%	68.1%	67.3%	67.7%	65.7%	65.8%	65.3%	65.3%
STRUCKCOLLISION	79.3%	79.2%	79.3%	79.1%	79.1%	79.2%	79.2%	79.0%	79.0%	79.1%	79.0%

Table D.11: Tables (F-measure) of ANOVA Test for Grouping Methods and Levene's Test

Source	DF	Sum of Squares	Mean Square	F Value	Pr >F
Model	106	1.15499345	0.01089616	11.93	<.0001
Error	2233	2.03945545	0.00091333		
Corrected Total	2339	3.1944489			

Source	DF	Type I SS	Mean Square	F Value	Pr >F
Category (block)	29	0.44063173	0.0151942	16.64	<.0001
Classifier	2	0.32652513	0.16326256	178.76	<.0001
Train-Test_Ratio	2	0.07442718	0.03721359	40.75	<.0001
Grouping_Method	9	0.13671929	0.01519103	16.63	<.0001
Classifier*Method	14	0.12839383	0.00917099	10.04	<.0001
Classifier*Ratio	4	0.01044439	0.0026111	2.86	0.0223
Method*Ratio	18	0.02350459	0.00130581	1.43	0.1073
Method*Ratio*Classifier	28	0.01434731	0.0005124	0.56	0.9694

Source	DF	Type III SS	Mean Square	F Value	Pr >F
Category (block)	29	0.44063173	0.0151942	16.64	<.0001
Classifier	2	0.26643816	0.13321908	145.86	<.0001
Train-Test_Ratio	2	0.07408074	0.03704037	40.56	<.0001
Grouping_Method	9	0.13671929	0.01519103	16.63	<.0001
Classifier*Method	14	0.12839383	0.00917099	10.04	<.0001
Classifier*Ratio	4	0.0062298	0.00155745	1.71	0.1461
Method*Ratio	18	0.02350459	0.00130581	1.43	0.1073
Method*Ratio*Classifier	28	0.01434731	0.0005124	0.56	0.9694

Levene's Test for Homogeneity of impact Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr >F**
Grouping_Method	9	0.00126	0.00014	6.97	<.0001
Error	2330	0.0467	0.00002		

Table D.12: Post Hoc Tests (Tukey and LSD) for Grouping Methods — MNB

Means with the same letter are not significantly different				
Tukey Grouping	**Mean**	**N**	**Grouping_Method**	**LSD Grouping**
A	**0.043877**	90	**MSS2PA**	A
A	0.042042	90	MSS2	A
A	0.037215	90	MSPA	A
A	0.035525	90	MS	A
B	**0.015755**	90	**STEM_DFC1**	B
B	**0.003481**	90	**LEMMA_DFC1**	C

Table D.13: Post Hoc Tests (Tukey and LSD) for Grouping Methods — SVM

Means with the same letter are not significantly different						
Tukey Grouping		**Mean**	**N**	**Grouping_Method**	**LSD Grouping**	
	A	**0.012164**	90	**MSS2PA**		A
	A	0.010324	90	MSPA	B	A
	A	0.005305	90	MSS2	B	C
B	A	0.0036	90	MS		C
B	A	**0.002875**	90	**STEM_DFC1**		C
B	A	**0.002822**	90	**LEMMA_DFC1**		C
B		-0.005524	90	MSS2PA_NB		D
B		-0.005886	90	MSPA_NB		D
	C	-0.017206	90	MSS2_NB		E
	C	-0.018502	90	MS_NB		E

Table D.14: Post Hoc Tests (Tukey and LSD) for Grouping Methods — LR

Means with the same letter are not significantly different								
Tukey Grouping			**Mean**	**N**	**Grouping_Method**	**LSD Grouping**		
	A		**0.026939**	90	**MSS2PA_NB**		A	
B	A		**0.025071**	90	**MSPA_NB**		A	
B	A		**0.022022**	90	**MSS2PA**		A	
B	A	C	0.019898	90	MSPA	B	A	
B	D	C	0.013664	90	MSS2_NB	B	C	
B	D	C	0.012113	90	MS_NB	B	C	D
	D	C	0.008816	90	MSS2		C	D
	D		0.006458	90	MS		C	D
	D		**0.006289**	90	**STEM_DFC1**		C	D
	D		**0.004915**	90	**LEMMA_DFC1**			D

VITA

VITA

Hsin-Ying Huang received her bachelor's degree in Industrial Engineering from National Chiao Tung University in 2008 and her master's degree in Industrial Engineering from National Tsing Hua University in 2010. She started her doctoral degree in Industrial Engineering at Purdue University in 2010. Her primary research interest is to explore, develop, and evaluate human-machine collaborative methods for improving statistical text analysis.